Contrasting Linguistic Patterns in Human and LLM-Generated News Text

Alberto Muñoz-Ortiz
alberto.munoz.ortiz@udc.es

University of A Coruña

Carlos Gómez-Rodríguez
University of A Coruña

David Vilares
University of A Coruña

Research Article

Keywords: Large Language Models, Computational Linguistics, Machine-generated text, Linguistic Biases

Posted Date: March 14th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4077382/v1

Additional Declarations: No competing interests reported.

Contrasting Linguistic Patterns in Human and LLM-Generated News Text

Alberto Muñoz-Ortiz[1*], Carlos Gómez-Rodríguez[1] and David Vilares[1]

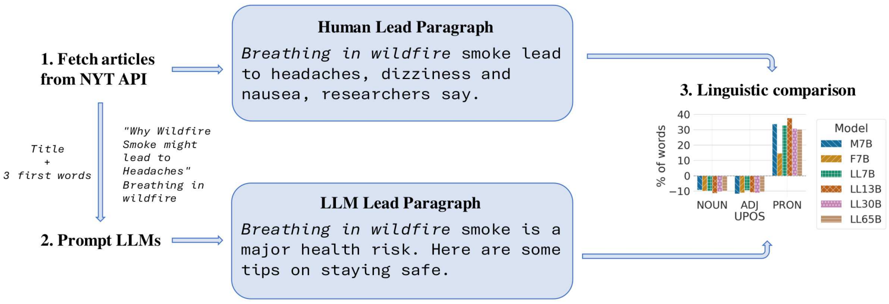
[1*]Departamento de Ciencias de la Computación y Tecnologías de la Información, Universidade da Coruña, CITIC, Campus de Elviña s/n, A Coruña, 15071, A Coruña, Spain.

*Corresponding author(s). E-mail(s): alberto.munoz.ortiz.com; Contributing authors: carlos.gomez@udc.es; david.vilares@udc.es;

Abstract

We conduct a quantitative analysis contrasting human-written English news text with comparable large language model (LLM) output from from six different LLMs that cover three different families and four sizes in total. Our analysis spans several measurable linguistic dimensions, including morphological, syntactic, psychometric, and sociolinguistic aspects. The results reveal various measurable differences between human and AI-generated texts. Human texts exhibit more scattered sentence length distributions, more variety of vocabulary, a distinct use of dependency and constituent types, shorter constituents, and more optimized dependency distances. Humans tend to exhibit stronger negative emotions (such as fear and disgust) and less joy compared to text generated by LLMs, with the toxicity of these models increasing as their size grows. LLM outputs use more numbers, symbols and auxiliaries (suggesting objective language) than human texts, as well as more pronouns. The sexist bias prevalent in human text is also expressed by LLMs, and even magnified in all of them but one. Differences between LLMs and humans are larger than between LLMs.

Keywords: Large Language Models, Computational Linguistics, Machine-generated text, Linguistic Biases



Human Lead Paragraph

Breathing in wildfire smoke lead to headaches, dizziness and nausea, researchers say.

LLM Lead Paragraph

Breathing in wildfire smoke is a major health risk. Here are some tips on staying safe.

Fig. 1: We gather contemporary articles from the New York Times API and use their headlines plus the 3 first words of the lead paragraph as prompts to LLMs to generate news. We use four LLMs from the LLaMa family (7B, 13B, 30B and 65B sizes), Falcon 7B and Mistral 7B. We then compare both types of texts, assessing differences in aspects like vocabulary, morphosyntactic structures, and semantic attributes

1 Introduction

Large language models (LLMs) (Radford et al., 2018; Scao et al., 2022; Touvron et al., 2023) and instruction-tuned variants (OpenAI, 2023; Taori et al., 2023) output fluent, human-like text in many languages, English being the best represented. The extent to which these models truly understand semantics (Landgrebe and Smith, 2021; Søgaard, 2022), encode representations of the world (Li et al., 2022), generate fake statements (Kumar et al., 2023), or propagate specific moral and ethical values (Santurkar et al., 2023) is currently under active debate. Regardless, a crucial factor contributing to the persuasiveness of these models lies, in the very first place, in their exceptional linguistic fluency.

A question that arises regards whether their storytelling strategies align with the linguistic patterns observed in human-generated texts. Do these models tend to use more flowery or redundant vocabulary? Do they exhibit preferences for specific voices or syntactic structures in sentence generation? Are they prone to certain psychometric dimensions? However, contrasting such linguistic patterns is not trivial. Firstly, the creators of these models often insufficiently document the training data used. Even with available information, determining the extent of the training set's influence on a sentence or whether it is similar to an input sample remains challenging. Second, language is subject to cultural norms, social factors, and geographic variations, which shape linguistic preferences and conventions. Thus, to contrast linguistic patterns between humans and machines, it is advisable to rely on a controlled environment. However, little effort has been made to measure differences, if any, in syntax, grammar, and other linguistic aspects between the two types of texts. Instead, attention has primarily been on explicit biases like societal and demographic biases (Liang et al., 2021).

Research contributions and objectives

We study six generative large language models: Mistral 7B (Jiang et al., 2023), Falcon 7B (Almazrouei et al., 2023) and the four models (7B, 13B, 30B and 65B) from the LLaMa family (Touvron et al., 2023). We contrast several linguistic patterns against human text using English news text. To do so, we recover human-generated news and ask the models to generate a news paragraph based on the headline and first words of the news. We query the New York Times Archive API to retrieve news published after all the models used were released, to guarantee sterilization from the training set. We analyze various linguistic patterns: differences in the distribution of the vocabulary, sentence length, part-of-speech (PoS) tags, syntactic structures, psychometric features such as the tone of the news articles and emotions detectable in the text, and sociolinguistic aspects like gender bias. We depict an overview in Figure 1. We also explore if these disparities change across models of different sizes and families.

2 Related work

Next, we survey relevant work to the subject of this paper: (i) analyzing inherent linguistic properties of machine-generated text, (ii) distinguishing between machine- and human-generated texts, (iii) applications of machine-generated text to economize human labor and time, (iv) generating synthetic text to perform data augmentation, and (v) examining the propagation of biases through data extracted from the Internet.

2.1 Analysis of linguistic properties of AI-generated text

Cognitive scientists (Cai et al., 2023) have exposed models such as ChatGPT to experiments initially designed for humans. They verified that it was able to replicate human patterns like associating unfamiliar words to meanings, denoising corrupted sentences, or reusing recent syntactic structures, among other abilities. Yet, they also showed that ChatGPT tends to refrain from using shorter words to compress meaning, as well as from using context to resolve syntactic ambiguities. Similarly, Zhou et al. (2023) conducted a thorough comparison between AI-created and human-created misinformation. They first curated a dataset of human-created misinformation pertaining to the COVID-19 pandemic. Then, they used these representative documents as prompts for GPT-3 to generate synthetic misinformation. By analyzing and contrasting the outputs from both sources, the study revealed notable differences. AI-made fake news tended to be more emotionally charged, using eye-catching language. It also frequently raised doubts without proper evidence and jumped to unfounded conclusions. Very recently, Xu et al. (2023) have shed light on the lexical conceptual representations of GPT-3.5 and GPT-4. Their study demonstrated that these AI language models exhibited strong correlations with human conceptual representations in specific dimensions, such as emotions and salience. However, they encountered challenges when dealing with concepts linked to perceptual and motor aspects, such as visual, gustatory, hand/arm, or mouth/throat aspects, among others. With the goal of measuring differences across both types of texts, Pillutla et al. (2021) introduced MAUVE, a new metric designed to compare the learned distribution of a language generation model with the distributions observed in human-generated texts. Given the inherent challenge in open-ended text generation, where there is no single correct output, they address the issue of gauging proximity between distributions by leveraging the concept of a divergence curve. Following the release of this work as a preprint, other authors have studied the text generated by language models from a linguistic point of view. Martínez et al. (2023) developed a tool to evaluate the vocabulary knowledge of language models, testing it on ChatGPT. Other works have also evaluated the lexical abundance of ChatGPT and how it varies with regards to different parameters (Martínez et al., 2024). Linguistic analysis is proving to be a valuable tool in understanding LLM outputs. In the line of our work, Rosenfeld and Lazebnik (2024) conducted a linguistic analysis of the outputs from three popular LLMs, concluding that this type of information can be used for LLM attribution on machine-generated texts.

2.2 Identification of synthetically-generated text

This research line aims to differentiate texts generated by machines from those authored by humans (Crothers et al., 2023), thus contributing to accountability and transparency in various domains. This challenge has been addressed from different angles including statistical, feature-based methods (Nguyen-Son et al., 2017; Fröhling and Zubiaga, 2021) and neural approaches (Rodriguez et al., 2022; Zhan et al., 2023). Yet, Crothers et al. (2022) recently concluded that except from neural methods, the other approaches have little capacity to identify modern machine-generated texts. Ippolito et al. (2020) observed two interesting behaviors related to this classification task: (i) that fancier sampling methods for generation (e.g., nucleus or untruncated random sampling) are helpful to better at deveiving humans, but conversely make the detection for machines more accessible and simple, and (ii) that showing longer inputs help both machines and humans to better detect synthetically-generated strings. Munir et al. (2021) showed that it was possible to attribute a given synthetically-generated text to the specific LLM model that produced it, using a standard machine learning classification architecture that used XLNet (Yang et al., 2019) as its backbone. In a different line, Dugan et al. (2020) studied whether humans could identify the fencepost where an initially human-generated text transitions to a machine-generated one. There are also methods that have been specifically designed to generate or detect machine-generated texts for highly sensible domains, warning about the dangers of language technologies. The SCIgen software (Stribling et al., 2005) was able to create semantically non-sense but grammatically correct research papers, whose content was accepted at some conferences with poor peer-review processes. More recently, Liao et al. (2023) showed that medical texts generated by ChatGPT were easy to detect: although the syntax is correct, the texts were more vague and provided only only general terminology or knowledge. However, this is a hard task and methods to detect AI-generated text are not accurate and are susceptible to suffer attacks (Sadasivan et al., 2023).

2.3 Natural language annotation and data generation using LLMs

The quality of current synthetically-generated text has encouraged researchers to explore their potential for complementing labor-intensive tasks, such as annotation and evaluation. For instance, He et al. (2022) generated synthetic unlabeled text tailored for a specific NLP task. Then, they used an existing supervised classifier to silver-annotate those sentences, aiming to establish a fully synthetic process for generating, annotating, and learning instances relevant to the target problem. Related, Chiang and Lee (2023) investigated whether LLMs can serve as a viable replacement for human evaluators in downstream tasks. Particularly, they conducted experiments where LLMs are prompted with the same instructions and samples as provided to humans, revealing a correlation between the ratings assigned by both types of evaluators. Moreover, there is also work to automatically detect challenging samples in datasets. For instance, Swayamdipta et al. (2020) already used the LLMs fine-tuning phase to identify simple, hard and ambiguous samples. Chong et al. (2022) demonstrated that language models are useful to detect label errors in datasets by simply ranking the loss of fine-tuned data.

LLMs can also contribute in generating high-quality texts to pretrain other models. Previous work has used language models to generate synthetic data to increase the amount of available data using pretrained models (Kumar et al., 2020). Some examples of downstream tasks are text classification (Li et al., 2023), intent classification (Sahu et al., 2022), toxic language detection (Hartvigsen et al., 2022), text mining (Tang et al., 2023), or mathematical reasoning (Liu et al., 2023b), *inter alia*. Synthetic data is also used to pretrain and distill language models. Data quality has been shown to be a determinant factor for training LLMs. Additional synthetic data can contribute to scale the dataset size to compensate a small model size, getting more capable small models. LLMs have allowed to generate high-quality, synthetic text that is useful to train small language models (SLM). One of such cases is (Eldan and Li, 2023). They generated high quality data with a constrained vocabulary and topics using GPT-3.5 and 4 to train SLM that show coherence, creativity and reasoning in a particular domain. The Phi models family (Gunasekar et al., 2023; Li et al., 2023a; Javaheripi et al., 2023) showed the usefulness of synthetic data in training high-performance but SLMs. The authors used a mixture of high-quality textbook data and synthetically-generated textbooks to train a highly-competent SLM. Moreover, it has been used to create instruction tuning datasets to adequate LLMs behavior to user prompts (Peng et al., 2023). Synthetic data can also help preventing LLMs from adapting their answers to previous human opinions when they are not objectively correct (Wei et al., 2023). However, although useful, synthetically-generated data may harm performance when the tasks or instances at hand are subjective (Li et al., 2023).

Synthetic datasets provide data whose content is more controllable, as LLMs tend to reproduce the structure of the datasets they have been trained on. Most LLMs are trained totally or partially on scraped data from the web, and such unfiltered internet data usually contain biases or discrimination as they reproduce the hegemonic view (Bender et al., 2021). Some widely-used huge datasets such as The Pile (Gao et al., 2020) confirm this. Authors extracted co-occurrences on the data that reflect racial, religious and gender stereotypes, which are also shown in some models. Some datasets are filtered and refined to improve the quality of the data. However, they still reproduce the biases in it (Penedo et al., 2023). Moreover, Dodge et al. (2021) did an extensive evaluation of the data of the C4 dataset (Raffel et al., 2020), pointing out filtering certain information could increase the bias on minorities. Prejudices on the data are reproduced on the LLMs trained on them, as some studies have pointed out (Weidinger et al., 2021). LLMs show the same biases that occur in the datasets, ranging from religious (Abid et al., 2021) to gender discrimination (Lucy and Bamman, 2021).

3 Data preparation

Next, we will delve into our data collection process for both human- and machine-generated content, before proceeding to the analysis and comparison.

3.1 Data

We generate the evaluation dataset relying on news published after the release date of the models that we will use in this work. This strategy ensures that they did not have exposure to the news headlines and their content during pre-training. It is also in line with strategies proposed by other authors - such as Liu et al. (2023) - who take an equivalent angle to evaluate LLMs in the context of generative search engines. The reference human-generated texts will be the news (lead paragraph) themselves.

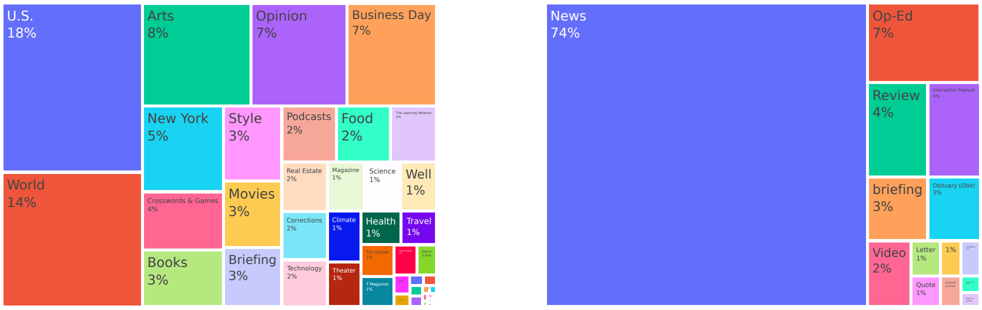
Crawling

We use New York Times news, which we access through its Archive API[1]. Particularly, we gathered all articles available between October 1, 2023, and January 24, 2024, resulting in a dataset of 13,371 articles. The articles are retrieved in JSON format, and include metadata such as the URL, section name, type of material, keywords, or publication date. Figure 2 shows some general information about the topics and type of articles retrieved. We are mainly interested on two fields: the headline and the lead paragraph. The lead paragraph is a summary of the information presented in the article. We discarded the articles that had an empty lead paragraph. The collected articles primarily consist of news pieces, although around 26% also include other types of texts, such as reviews, editorials or obituaries.

Rationale for focusing on English and the news domain

The choice to focus on English texts in our research is guided by a couple of considerations. Firstly, the LLMs we use (as detailed in Section 3.2) are English-centric. LLaMa's dataset comprises over 70% English content, and Falcon's even higher at over 80%. With Mistral, the specifics of the training data were not disclosed, adding an extra layer of complexity. In this context, it is worth noting that a model trained predominantly on data from specific demographics or regions might develop a bias towards those linguistic patterns, potentially overlooking others. The clarity around

[1] https://developer.nytimes.com/docs/archive-product/1/overview



(a) Section name

(b) Type of material

Fig. 2: Treemaps for the 'section name' and 'type of material' fields of the crawled articles

the influence of diverse linguistic inputs on model performance is also limited, further complicating a fair analysis.

Secondly, studying additional languages presented extra logistical challenges. Collecting non-English news texts was difficult, especially with the reliance on paid or services with very limited capabilties for access to quality sources. The abundance and accessibility of English news sources, like the New York Times, greatly facilitated our collection of analyzable content under usable licenses.

3.2 Generation

Let $\mathcal{H} = [h_1, h_2, ..., h_N]$ be a set of human-generated texts, such that h_i is a tuple of the form (t_i, s_i) where t_i is a headline and s_i is a paragraph of text with a summary of the corresponding news. Similarly, we will define $\mathcal{M} = [m_1, m_2, ..., m_N]$ as the set of machine-generated news articles produced by a LLM such that m_i is also a tuple of the from (t'_i, s'_i) where $t'_i = t_i$ and $s'_i = [w'_1, w'_2, ..., w'_{|s_i|}]$ is a piece of synthetic text. For the generation of high-quality text, language models aim to maximize the probability of the next word based on the previous content. To ensure that the models keep on track with the domain and topic, we initialize the previous content with the headline (the one chosen by the journalist that released the news) and the first three words of the human-generated lead paragraph to help the model start and follow the topic.[2] Formally, we first condition the model on $c_i = t'_i \cdot s_{i[0:2]}$ and every next word $(i \geq 3)$ will be predicted from a conditional distribution $P(w'_i | c_i \cdot s'_{i[3:t-1]})$.

To generate a piece of synthetic text s', we condition the models with a prompt that includes the headline and first words, as described above, and we keep generating news text until the model decides to stop.[3] We enable the model to output text

[2] In preliminary experiments, certain LLM outputs encountered difficulties in adhering to a minimal coherent structure when a minimum number of the body's words were absent from the prompt. Also note that the LLMs we are using are not instruction-tuned, and thus prompting engineering is not particularly suitable, nor the goal of this work.

[3] In preliminary experiments, we explored hyperparameter values that generated fluent and coherent texts: temperature of 0.7, 0.9 top p tokens, and a repetition penalty of 1.1.

without any forced criteria, except for not exceeding 200 tokens. The length limit serves two main purposes: (i) to manage computational resources efficiently[4], and (ii) to ensure that the generated content resembles the typical length of human-written lead paragraphs, making it comparable to human-produced content. We arrived at this limit after comparing the average and standard deviation of the number of tokens between humans and models in early experiments.

3.3 Selected models

We rely on six pre-trained generative language models that are representative within the NLP community. These models cover 4 different sizes (7, 13, 30 and 65 billion parameters) and 3 model families. We only include different sizes for LLaMa as results within the same family are similar, and larger models need considerably more compute. We briefly mention their main particularities below:

LLaMa models (LL) (Touvron et al., 2023)

The main representative for our experiments will be the four models from the LLaMa family, i.e. the 7B, 13B, 30B, and 65B models. The LLaMa models are trained on a diverse mix of data sources and domains, predominantly in English, as detailed in Table 1. LLaMa is based on the Transformer architecture and integrates several innovations from other large language models. In comparison to larger models like GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2023), and Chinchilla (Hoffmann et al., 2022), LLaMa exhibits superior performance in zero and few-shot scenarios. It is also a good choice as a representative example because the various versions, each with a different size, will enable us to examine whether certain linguistic patterns become closer or more different to humans in larger models.

Falcon 7B (F7B) (Almazrouei et al., 2023)

Introduced alongside its larger variants with 40 and 180 billion parameters, it is trained on 1.5 trillion tokens from a mix of curated and web datasets (see Table 1). Its architecture relies on multigroup attention (an advanced form of multiquery attention), Rotary Embeddings (similar to LLaMa), standard GeLU activation, parallel attention, MLP blocks, and omits biases in linear layers. We primarily chose this model to compare the results in the following sections with those of its counterpart, LLaMa 7B, and to explore whether there are significant differences among models of similar size.

Mistral 7B (M7B) (Jiang et al., 2023)

It surpasses larger LLaMa models in various benchmarks despite its smaller size. Its distinctive architecture features Sliding Window Attention, Rolling Buffer Cache, and Prefill and Chunking. The training data for Mistral 7B is not publicly disclosed, and to fight against data contamination issues, our analysis only includes articles published after the model's release. The choice of this model as an object of study follows the

[4] We ran the models on 2xA100 GPUs for 3 days to generate all texts. To address memory costs, we use 8-bit precision.

Table 1: Size and training data of the models used in our experiments

Family	Size	Tokens	Data sources
LLaMa	7B 13B 30B 65B	1T 1T 1.5T 1.5T	English CommonCrawl (67%), C4 (15%), GitHub (4.5%), Wikipedia (4.5%), Gutenberg and Books3 (4.5%), ArXiv (2.5%), Stack Exchange (2%)
Falcon	7B	1.5T	RefinedWeb-English (76%), RefinedWeb-Euro (8%), Gutenberg (6%), Conversations (5%) GitHub (3%), Technical (2%)
Mistral	7B	?	?

same thinking we used for the Falcon model. We want to see how well Mistral 7B does and how its new features stack up against models of the same size.

4 Analysis of linguistic patterns

In this section, we compare human- and machine-generated texts. We first inspect the texts under a morphosyntactic optic, and then focus on semantic aspects.

4.1 Morphosyntactic Analysis

To compute linguistic representations, we rely on Stanza (Qi et al., 2020) to perform segmentation, tokenization, part-of-speech (PoS) tagging, and dependency and constituent parsing. For these tasks, and in particular for the case of English and news text, the performance is high enough to be used for applications (Manning, 2011; Berzak et al., 2016), and it can be even superior to that obtained by human annotations. This also served as an additional reason to focus our analysis on news text, ensuring that the tools we rely on are accurate enough to obtain meaningful results.

4.1.1 Sentence length

Figure 3 illustrates the length distribution for the LLMs in comparison to human-generated news articles. We excluded a few outliers from the plot by ignoring sentences with lengths over 80 tokens. The six LLMs exhibit a similar distribution across different sentence lengths, presenting less variation when compared to human-generated sentences, which display a wider range of lengths and greater diversity. Specifically, the models exhibit a higher frequency of sentence generation within the 10 to 30 token range compared to humans, whereas humans tend to produce longer sentences with greater frequency.

4.1.2 Richness of vocabulary and lexical variation

We analyze the diversity of vocabulary used by the LLMs and compare them against human texts. We consider the total number of tokens (words), the count of unique tokens, and the Type-Token Ratio (TTR). The TTR is a measure of lexical variation and is calculated by dividing the number of types (i.e., unique tokens) by the total

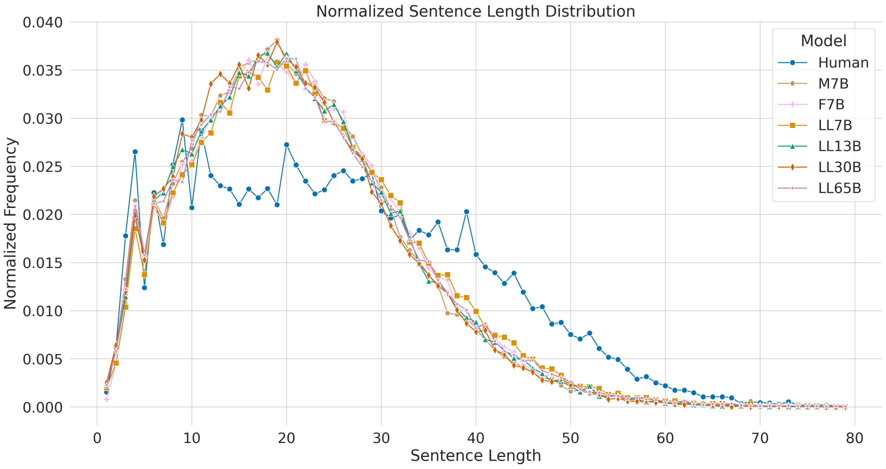


Fig. 3: Sentence length distribution for the human-written texts and each tested language model. M stands for Mistral, F for Falcon and LL for LLaMa

number of tokens. Table 2 presents the results. Human texts displayed a higher diversity of unique tokens compared to most LLMs — except for the 65B LLaMa model — despite using fewer total tokens. This resulted in a higher TTR and a richer vocabulary.

In terms of model size, we see that all LLaMa versions have a similar number of unique tokens and TTR, but there is a moderate increasing trend as model size increases. Comparing models of the same size (LLaMa 7B, Falcon 7B, and Mistral 7B), Falcon 7B is the model that uses the fewest unique tokens (in absolute terms) by a wide margin. However, it shows the same TTR as LLaMa 65B, as it also tends to generate shorter texts and thus the ratio is similar. Finally, Mistral 7B has the lowest TTR and the second lowest number of unique tokens.

Table 2: Statistics related to the vocabulary of the articles generated by humans and each tested language model

Model	**Tokens**	**Unique**	**Type-token ratio**
Human	676 591	39 058	0.058
M7B	741 489	34 399	0.041
F7B	606 020	29 262	0.048
LL7B	843 087	37 553	0.045
LL13B	809 551	37 091	0.046
LL30B	790 059	38 390	0.049
LL65B	824 739	39 881	0.048

Table 3: UPOS frequencies (%) in human- and LLM-generated texts

UPOS	H	M7B	F7B	LL7B	LL13B	LL30B	LL65B
NOUN	19.69	17.85	17.72	17.75	17.44	17.64	17.74
PUNCT	11.88	10.92	12.14	10.77	10.91	11.43	11.22
ADP	11.36	10.58	10.30	10.75	10.63	10.70	10.69
VERB	9.97	10.37	9.23	10.26	10.23	10.14	10.29
PROPN	9.61	8.75	9.44	9.14	9.18	9.52	9.50
DET	9.04	9.00	10.72	8.65	8.64	8.76	8.63
ADJ	7.58	6.69	6.74	6.86	6.76	6.73	6.77
PRON	5.32	7.12	6.11	7.08	7.33	6.96	6.93
AUX	3.81	5.77	6.02	5.65	5.74	5.50	5.41
ADV	3.26	3.41	2.61	3.58	3.68	3.41	3.49
CCONJ	2.65	2.72	2.52	2.68	2.70	2.61	2.67
PART	2.43	2.76	2.80	2.64	2.63	2.52	2.58
NUM	1.77	1.95	1.98	2.02	1.98	2.05	2.02
SCONJ	1.41	1.84	1.37	1.84	1.85	1.71	1.72
INTJ	0.12	0.08	0.08	0.08	0.08	0.08	0.09
SYM	0.09	0.17	0.19	0.19	0.19	0.18	0.18
X	0.03	0.03	0.02	0.05	0.04	0.06	0.07

4.1.3 Part-of-speech tag distributions

Table 3 presents the frequency of universal part-of-speech (UPOS) tags (Petrov et al., 2012) for both human and LLM-generated texts. Figure 4 shows relative differences observed across humans and each model, for a better understanding of the relative use of certain grammatical categories. Overall, the behavior of LLMs and their generated text tends to be consistent among themselves, yet shows differences when compared to human behavior, i.e., they exhibit in some cases a greater or lesser use of certain grammatical categories. To name a few, humans exhibit a preference for using certain kinds of content words, such as nouns and adjectives. Humans also use words tagged as punctuation symbols more often (except when compared to Falcon), which may be connected to sentence length, as human users tend to rely on longer sentences, requiring more punctuation. Alternatively, the language models exhibit a pronounced inclination towards relying on categories such as symbols or numbers, possibly indicating an extra effort by language models to furnish specific data in order to sound convincing. Moreover, they write pronouns more frequently; we will analyze this point later from a gender perspective. Comparing LLM families, Mistral and LLaMa show a similar use of POS tags, with Mistral being the model that resembles humans the most. Falcon, however, has some strong anomalies in POS tags such as DET or ADV. Regarding model size, the larger the model, the greater the similarity with humans. Nevertheless, differences between differently-sized models are much smaller than between models and humans.

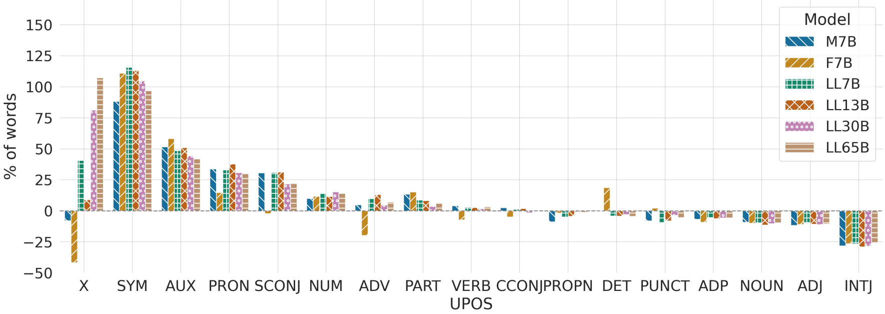


Fig. 4: Percentage differences, following Table 3, in the use of each UPOS category for each tested language model in comparison to humans

4.1.4 Dependencies

Dependency arc lengths

Table 4 shows information about the syntactic dependency arcs in human and machine-generated texts. In this analysis, we bin sentences by length intervals to alleviate the noise from comparing dependency lengths on sentences of mixed lengths (Ferrer-i-Cancho and Liu, 2014). Results indicate that dependency lengths and their distributions are nearly identical for all the LLMs except Falcon and the human texts, which both tend to use longer dependencies than the texts by the rest of the LLMs. This finding holds true for every sentence length bin for Falcon, and for all but the first (length 1-10) in the case of human texts, so we can be reasonably sure that it is orthogonal to the variation in sentence length distribution between human and LLM texts described earlier. It is also worth noting that, in spite of the similarities between humans and Falcon in terms dependency lengths, their syntax is not that similar overall: there is a substantial difference in directionality of dependencies, with Falcon using more leftward dependencies than both humans and other LLMs. The fact that Falcon-generated texts are not really human-like in terms of dependency syntax is further highlighted in the next section, where we consider a metric that normalizes dependency lengths.

Optimality of dependencies

We compare the degree of optimality of syntactic dependencies between human texts and LLMs. It has been observed in human language that dependencies tend to be much shorter than expected by chance, a phenomenon known as dependency length minimization (Ferrer-i-Cancho, 2004; Futrell et al., 2015). This can be quantified in a robust way (with respect to sentence length, tree topology and other factors) by the Ω optimality score introduced in Ferrer-i Cancho et al. (2022). This score measures where observed dependency lengths sit with respect to random word orders and optimal word orders, and is defined as: $\Omega = \frac{D_{rla} - D}{D_{rla} - D_{min}}$, where D is the sum of dependency lengths in the sentence, D_{rla} is the expected sum of lengths, and D_{min} is the optimal sum of lengths for the sentence's tree structure. For optimally-arranged trees $D = D_{min}$ and

Table 4: Statistics for dependency arcs in sentences of different lengths for the texts generated by human writers and each tested language model. The meaning of the columns is as follows: (%L, %R) percentage of left and right arcs, ($\bar{l}$) average arc length, ($\bar{l}_L$, $\bar{l}_R$) average left and right arc length, (σ_l) standard deviation of arc length, (σ_{l_L}, σ_{l_R}) standard deviation of left and right arc length, and number of sentences

l	Model	%L	%R	$\bar{\mathbf{l}}$	$\bar{\mathbf{l}}_\mathbf{L}$	$\bar{\mathbf{l}}_\mathbf{R}$	$\sigma_\mathbf{l}$	$\sigma_{\mathbf{l}_\mathbf{L}}$	$\sigma_{\mathbf{l}_\mathbf{R}}$	# Sent
1-10	Human	49.40	50.60	2.37	2.89	1.84	1.67	1.90	1.17	4 719
	M7B	50.94	49.06	2.37	2.93	1.83	1.65	1.88	1.16	6 190
	F7B	52.08	47.92	2.39	2.99	1.84	1.62	1.84	1.15	4 596
	LL7B	50.68	49.32	2.37	2.95	1.81	1.65	1.88	1.14	6 114
	LL13B	50.42	49.58	2.37	2.94	1.81	1.65	1.88	1.14	6 711
	LL30B	49.97	50.03	2.37	2.92	1.81	1.65	1.89	1.14	6 808
	LL65B	50.23	49.77	2.36	2.91	1.81	1.64	1.87	1.15	6 652
11-20	Human	58.36	41.64	3.19	4.62	2.17	3.12	3.87	1.86	6 179
	M7B	59.76	40.24	3.12	4.63	2.10	3.03	3.80	1.74	12 113
	F7B	61.41	38.59	3.20	4.79	2.19	3.06	3.85	1.83	9 265
	LL7B	59.74	40.26	3.11	4.63	2.09	3.03	3.81	1.72	12 361
	LL13B	59.69	40.31	3.12	4.63	2.11	3.03	3.81	1.75	12 762
	LL30B	59.62	40.38	3.12	4.63	2.11	3.03	3.80	1.76	13 039
	LL65B	59.43	40.57	3.13	4.63	2.10	3.04	3.81	1.75	12 767
21-30	Human	60.40	39.60	3.64	5.52	2.41	4.42	5.71	2.68	6 153
	M7B	61.00	39.00	3.53	5.50	2.26	4.28	5.65	2.33	10 449
	F7B	62.51	37.49	3.62	5.70	2.38	4.32	5.72	2.46	8 222
	LL7B	60.87	39.13	3.51	5.47	2.25	4.26	5.64	2.30	11 014
	LL13B	60.86	39.14	3.53	5.49	2.27	4.27	5.64	2.34	11 017
	LL30B	60.71	39.29	3.53	5.48	2.27	4.26	5.61	2.34	10 810
	LL65B	60.47	39.53	3.53	5.47	2.26	4.28	5.63	2.35	10 884
31-40	Human	60.84	39.16	3.90	6.07	2.50	5.49	7.32	3.19	4 770
	M7B	60.48	39.52	3.79	5.95	2.38	5.35	7.15	2.98	4 676
	F7B	61.98	38.02	3.89	6.11	2.52	5.35	7.16	3.12	4 064
	LL7B	60.79	39.21	3.78	5.98	2.35	5.34	7.19	2.90	5 790
	LL13B	60.51	39.49	3.79	5.96	2.38	5.33	7.14	2.93	5 280
	LL30B	60.35	39.65	3.81	5.95	2.40	5.33	7.09	2.99	4 949
	LL65B	60.35	39.65	3.79	5.95	2.37	5.31	7.10	2.93	5 430
+41	Human	60.48	39.52	4.01	6.28	2.53	6.20	8.32	3.58	2 967
	M7B	60.09	39.91	3.95	6.23	2.44	6.24	8.45	3.39	1 415
	F7B	61.77	38.23	4.04	6.43	2.56	6.18	8.46	3.44	1 318
	LL7B	59.83	40.17	3.97	6.25	2.44	6.24	8.41	3.43	2 035
	LL13B	60.47	39.53	3.99	6.29	2.48	6.23	8.41	3.50	1 693
	LL30B	60.21	39.79	3.98	6.24	2.49	6.21	8.34	3.53	1 579
	LL65B	60.08	39.92	3.95	6.22	2.45	6.16	8.33	3.37	1 880

Ω takes a value of 1, whereas for a random arrangement it has an expected value of 0. Negative values are possible (albeit uncommon) if dependency lengths are larger than expected by chance.

Figure 5 displays the distribution of Ω values across sentences for human and LLM-generated texts. The values were calculated using the LAL library (Alemany-Puig et al., 2021). Results indicate that the distribution of Ω values is almost identical between all of the LLMs, but human texts show noticeably larger values. This means human texts are more optimized in terms of dependency lengths, i.e. they have shorter dependencies than expected by a larger margin than those generated by the LLMs. At a first glance, this might seem contradictory with the results in the previous section, which showed that human texts had *longer* dependencies on average than non-Falcon

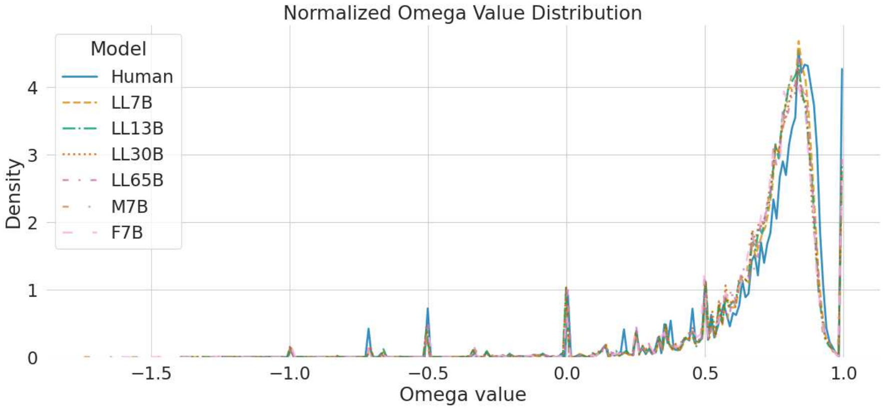


Fig. 5: Ω value distribution for the human- and LLM-generated texts

LLM texts. However, there is no real contradiction as the object of measurement is different, and in fact this is precisely the point of using Ω to refine and complement the previous analysis. While previously we measured dependency distances in absolute terms, Ω measures them controlling for tree topology, i.e., given the shape of a tree (e.g. a linear tree which is arranged as a chain of dependents, or a star tree where one node has all the others as dependents), are the words arranged in an order that minimizes dependencies *within possible with that shape?* Thus, combining the results from both sections we can conclude that while humans produce longer dependencies, this is due to using syntactic structures with different topology, but their word order is actually *more* optimized to make dependencies as short as possible. In turn, we also note that while Falcon's dependency lengths seemed different from the other LLMs (and more human-like) in absolute terms, the differences vanish (with all LLMs including Falcon having almost identical distributions, and humans being the outlier) when considering Ω.

Dependency types

Table 5 lists the frequencies for the main syntactic dependency types in human and machine-generated texts. We observe similar trends to the previous sections, with LLM texts exhibiting similar uses of syntactic dependencies among themselves, with Falcon being the most distinct model, while all of them present differences compared to human-written news. In terms of the LLaMa models - same model in different sizes - larger models are slightly closer to the way humans use dependency types. For the full picture, Figure 6 depicts all relative differences in their use (humans versus each LLM), but we briefly comment on a few relevant cases as representative examples. For instance, `nummod` dependencies are more common in LLM-generated texts compared to human texts. This is coherent with the higher use of the `NUM` tag in the part-of-speech tag distribution analysis. Additionally, we observed higher ratios for other dependency types, such as `aux` (for which the use of auxiliary verbs was also significantly higher according to the UPOS analysis), copula and nominal subjects (`nsubj`). Furthermore,

Table 5: Percentage of words generated by humans and each of the tested LLMs that are labeled with a specific dependency type (deprel). We only include relations with a frequency surpassing 1% within the human texts

deprel	H	M7B	F7B	LL7B	LL13B	LL30B	LL65B
punct	11.88	10.92	12.15	10.78	10.91	11.44	11.23
case	11.69	10.81	10.75	10.98	10.76	10.89	10.85
det	8.88	8.81	10.59	8.45	8.43	8.56	8.43
amod	6.98	5.57	5.73	5.79	5.60	5.71	5.75
nsubj	6.09	7.20	6.89	7.00	7.21	7.11	7.02
obl	5.50	5.24	4.67	5.39	5.31	5.36	5.31
nmod	4.95	4.45	4.84	4.50	4.40	4.47	4.47
compound	4.87	4.04	4.46	4.20	4.13	4.27	4.33
obj	4.28	4.41	3.91	4.22	4.23	4.19	4.27
advmod	3.46	3.63	2.91	3.83	3.98	3.65	3.76
conj	3.07	2.80	2.71	2.83	2.79	2.73	2.83
mark	2.65	3.35	2.94	3.27	3.28	3.07	3.12
cc	2.63	2.73	2.54	2.72	2.73	2.63	2.69
nmod:poss	2.34	2.21	2.01	2.21	2.19	2.19	2.17
flat	2.04	1.67	1.72	1.79	1.80	1.92	1.91
aux	1.91	2.74	2.72	2.68	2.71	2.58	2.55
advcl	1.80	1.67	1.30	1.69	1.70	1.62	1.67
cop	1.26	1.98	2.28	1.90	2.02	1.92	1.86
acl:relcl	1.22	1.33	1.29	1.38	1.29	1.26	1.28
appos	1.19	0.85	1.07	0.92	0.92	0.99	1.00
nummod	1.14	1.16	1.16	1.22	1.21	1.23	1.21
xcomp	1.10	1.40	1.27	1.37	1.36	1.30	1.34
acl	1.06	0.93	0.84	0.92	0.87	0.88	0.93

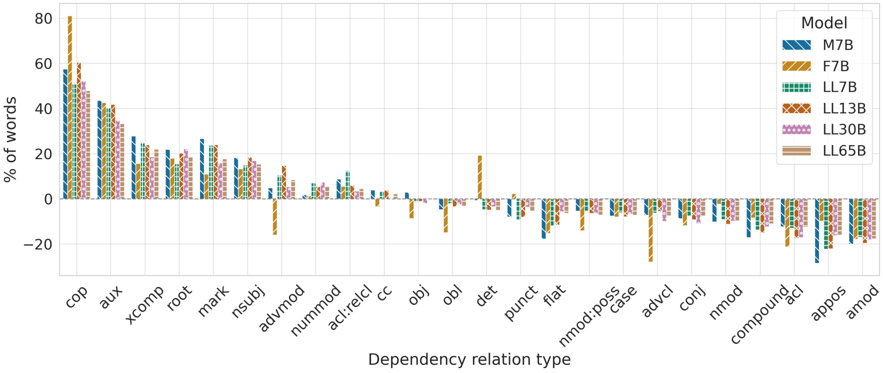


Fig. 6: Percentage differences, following Table 5, in the use of dependency relations for each tested language model in comparison to humans

syntactic structures from LLMs exhibit significantly fewer subtrees involving adjective modifiers (`amod` dependency type) and appositional modifiers (`appos`).

Table 6: Statistics for constituents that arise in sentences of different lengths for the text generated by human writers and each tested LLM. The meaning of the rows are: ($\bar{l}$) average constituent length, (σ_l) standard deviation of constituent length, and number of sentences

	Model	**1-10**	**11-20**	**21-30**	**31-40**	**+41**
$\bar{l}$	H	4.32	6.37	7.90	9.38	10.60
	M7B	4.39	6.55	8.27	9.77	11.01
	F7B	4.43	6.47	8.03	9.47	10.76
	LL7B	4.40	6.57	8.33	9.89	11.19
	LL13B	4.40	6.55	8.27	9.76	11.01
	LL30B	4.40	6.49	8.21	9.68	10.86
	LL65B	4.36	6.53	8.25	9.73	10.96
σ_l	H	2.35	4.64	6.97	9.19	11.24
	M7B	2.35	4.66	6.92	9.13	11.18
	F7B	2.33	4.63	6.80	8.94	11.01
	LL7B	2.35	4.69	6.99	9.24	11.35
	LL13B	2.33	4.68	6.95	9.14	11.19
	LL30B	2.36	4.66	6.94	9.14	11.17
	LL65B	2.34	4.68	6.96	9.17	11.23
# Sent	H	4 679	6 180	6 154	4 770	2 966
	M7B	6 108	12 113	10 448	4 678	1 414
	F7B	4 575	9 266	8 211	4 011	1 318
	LL7B	6 039	12 362	11 014	5 789	2 035
	LL13B	6 627	12 762	11 018	5 279	1 693
	LL30B	6 713	13 044	10 806	4 949	1 579
	LL65B	6 569	12 765	10 844	5 430	1 880

4.1.5 Constituents

Constituent lengths

Table 6 shows the comparison between the distribution of syntactic constituent lengths across both types of texts. While human-generated sentences, on average, surpass the length of those generated by LLMs, the average length of a sentence constituent for LLMs is observed to be greater than for humans. The standard deviation exhibits similar values across all models for each sentence length range. Similar to previous sections, Falcon 7B also displays the largest differences among language models. Within the LLaMa models, we can observe a clear decreasing trend with size which is broken by the 65B model, for which constituent lengths increase again across most of the length bins.

Table 7: Percentage of spans generated by humans and LLMs labeled with a specific constituent type. Only constituent types that conform more than 1% of the human's texts spans are shown

Type	H	M7B	F7B	LL7B	LL13B	LL30B	LL65B
NP	42.91	40.17	40.02	40.69	40.54	39.96	41.42
VP	18.08	20.18	20.29	19.97	20.02	20.59	20.19
PP	14.12	12.91	12.69	12.94	12.81	12.62	12.81
S	11.79	13.09	13.31	13.12	13.12	13.40	13.27
SBAR	3.64	4.34	4.29	4.09	4.15	4.34	3.84
ADVP	2.39	2.50	2.62	2.44	2.49	2.37	1.86
ADJP	1.97	1.78	1.82	1.76	1.79	1.75	1.80
NML	1.73	1.43	1.43	1.47	1.52	1.40	1.66
WHNP	1.41	1.47	1.49	1.53	1.44	1.46	1.47

Constituent types

Table 7 and Figure 7 examine the disparities in constituent types between human- and LLM-generated texts. Our focus was on constituent types that occur more than 1% of the times. Comparing humans and LLMs, some outcomes are in the same line of earlier findings: human-generated content displays heightened use of noun, adjective, and prepositional phrases (NP, ADJP, and PP, respectively). On the contrary, there is minimal divergence in the frequency of adverb phrases (ADVP) except for Falcon 7B, which shows a great difference. human and LLM-generated texts, the latter exhibits a more pronounced propensity for verb phrases (VP). Despite the similar frequency of the VERB UPOS tag in human and LLM-generated texts, the latter exhibit a more pronounced propensity for verb phrases (VP), consistent with the increased use of auxiliary verbs (whose UPOS tag is AUX, not VERB) that we saw in previous sections. Finally, we see that language models use a considerably larger amount of subordinate clauses (SBAR). Regarding model families, results are similar to those of dependencies and POS tags, but when looking at model size, previous trends are less obvious.

4.2 Semantic Analysis

As in the previous section, we are relying on state-of-the-art NLP models to accurately analyze different semantic dimensions: (i) emotions, (ii) text similarities, and (iii) gender biases, in an automated way.

4.2.1 Emotions

To study differences in the emotions conveyed by human- and LLM-generated outputs, we relied on the Hartmann (2022) emotion model. Table 8 provides the percentage of articles labeled with distinct emotional categories, including anger, disgust, fear, joy, sadness, surprise, and a special tag neutral to denote that no emotion is present in the text. Figure 8 depicts the percentage of articles associated with each emotion for each large language model used, as compared to human-written texts. As anticipated in journalistic texts, a substantial majority of the lead paragraphs are

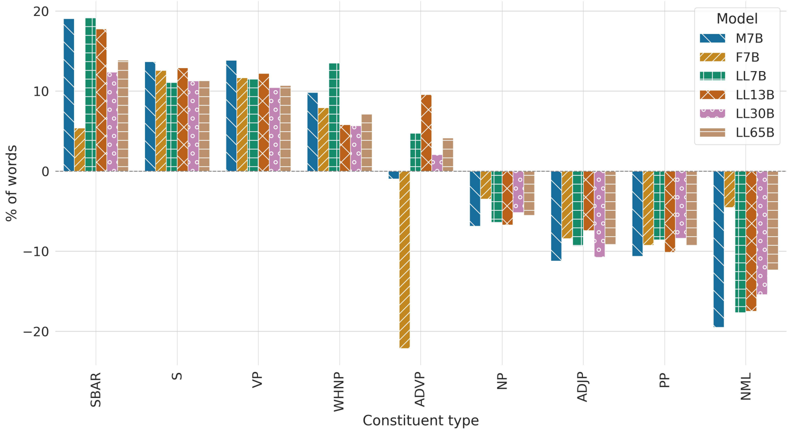


Fig. 7: Percentage differences, following Table 7, in the use of constituent labels for each tested language model in comparison to humans

classified as neutral. This category accounts for over 50% of the texts across all models and human-generated samples, with the LLM-generated text demonstrating a slightly higher inclination towards neutrality.

Concerning the remainder of the samples, human texts demonstrate a greater inclination towards negative and aggressive emotions like disgust and fear. However, humans and LLMs generated roughly the same amount of angry texts. In contrast, LLMs tend to generate more texts imbued with positive emotions, such as surprise and especially joy. The LLMs also produce many sad texts, a passive but negative emotion, yet less toxic than emotions such as anger or fear. Across LLaMa models, fear increases as the number of parameters grows (from LLaMa 13B), making them more akin to human texts. Since LLaMa (version 1 models) were not fine-tuned with reinforcement learning with human feedback, we hypothesize the main source contributing to this issue might be some pre-processing steps used for the LLaMa models, such as removing toxic content from its data. Yet, LLaMa's technical report (Touvron et al., 2023) mentioned an increase in model toxicity as they scaled up in size despite using the same pre-processing in all cases, which is coherent with our findings. When looking at families, Mistral comes closest to expressing emotions in a way similar to humans, and Falcon expresses more joy and less anger and surprise than the rest of the models.

4.2.2 Text similarity

We conducted an analysis of the cosine semantic similarity between lead paragraphs generated by various LLMs and their human-authored counterparts. Our objective was to investigate the impact of model sizes on the semantic similarity between these texts. To achieve so, we used a a state-of-the-art sentence similarity model called `all-mpnet-base-v2`[5] (Reimers and Gurevych, 2019). Figure 9 illustrates the distribution of the similarity scores obtained from our analysis. Results show that smaller-sized

[5] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Table 8: Percentage of articles generated by humans and LLMs that are labeled with different emotions

Model	Emotion						
	`anger`	`disgust`	`fear`	`joy`	`neutral`	`sadness`	`surprise`
H	8.04	9.35	10.77	8.30	52.16	8.51	2.87
M7B	7.29	7.65	8.34	9.80	53.83	9.72	3.37
F7B	6.11	8.32	8.77	8.53	56.55	8.99	2.73
LL7B	7.13	7.19	8.68	8.97	55.57	9.43	3.01
LL13B	7.72	7.41	8.69	9.00	53.95	9.72	3.51
LL30B	7.39	7.45	8.61	9.54	54.23	9.59	3.19
LL65B	7.45	8.26	9.25	9.10	53.65	8.80	3.49

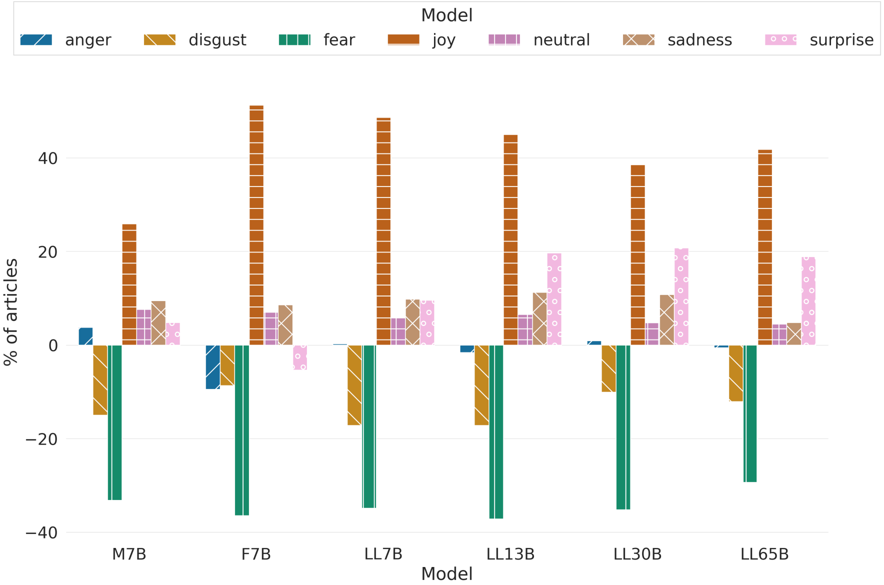


Fig. 8: Relative difference of emotion labels of articles generated by different LLMs in comparison to human texts

LLMs do not necessarily result in a decrease in sentence similarity compared to the human-authored texts. Differences across families are negligible.

4.2.3 Gender bias

Although related as well with our study with part-of-speech tag distribution, we here separately analyze the proportion between masculine and feminine pronouns used in both human- and LLM-generated text. Based on the morphological output by Stanza, we find the words that are pronouns have the features `Gender=Masc` and `Gender=Fem`, respectively. Results in Table 9 indicate that the already biased human texts use male

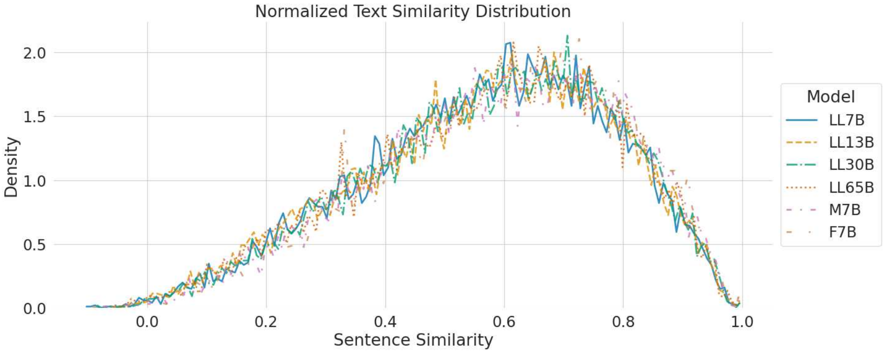


Fig. 9: Similarity scores between the sentences generated by the LLMs and human text

Table 9: Male-to-female ratio of pronouns used by the text generated by humans and each LLM

Model	Male-Female ratio	Difference with humans
H	1.71	-
M7B	1.74	3.06 %
F7B	1.64	-7.54 %
LL7B	1.86	14.30%
LL13B	1.89	17.13 %
LL30B	1.87	15.73 %
LL65B	1.88	17.04 %

pronouns 1.71 times more frequently than female pronouns. This is exacerbated by all models but Falcon 7B, which, although still heavily biased towards male pronouns, reduces the bias by 7.5%. LlaMa models, on the contrary, use around 15% more male than female pronouns in comparison to humans. This quantity is roughly the same for every size. Mistral 7B lies in the middle, with a slight increase of the male-female ratio of 3% with regards to human text.

5 Conclusion

This paper presented a comprehensive study on linguistic patterns in texts produced by both humans and machines, comparing them under controlled conditions. To keep up with current trends, we used modern generative models. To ensure the novelty of texts and address memorization concerns, we fed the LLMs headlines from news articles published after the release date of the models. The study revealed that despite generating highly fluent text, these models still exhibited noticeable differences when compared to human-generated texts. More precisely, at the lexical level, large language models relied on a more restricted vocabulary, except for LLaMa 65B. Additionally, at the morphosyntactic level, discernible distinctions were observed between human and machine-generated texts, the latter having a preference for parts of speech displaying (a sense of) objectivity - such as symbols or numbers - while using substantially fewer adjectives. We also observed variations in terms of syntactic structures, both for dependency and constituent representations, specifically in the use of dependency and constituent types, as well as the length of spans across both types of texts. In this respect our comparison shows, among other aspects, that all tested LLMs choose word orders that optimize dependency lengths to a lesser extent than humans; while they have a tendency to use more auxiliary verbs and verb phrases and less noun and prepositional phrases. In terms of semantics, while exhibiting a great text similarity with respect to the human texts, the models tested manifested less propensity than humans for displaying aggressive negative emotions, such as fear or anger. Mistral 7B generated texts whose emotion distributions are more similar to humans than those of LLaMa and Falcon models. However, we noted a rise in the volume of negative emotions with the models' size. This aligns with prior findings that associate larger sizes with heightened toxicity (Touvron et al., 2023). Finally, we detected an inclination towards the use of male pronouns, surpassing the frequency in comparison to their human counterparts. All models except Falcon 7B exacerbated this bias.

Funding

We acknowledge the European Research Council (ERC), which has funded this research under the Horizon Europe research and innovation programme (SALSA, grant agreement No 101100615); SCANNER-UDC (PID2020-113230RB-C21) funded by MICIU/AEI/10.13039/501100011033; Xunta de Galicia (ED431C 2020/11); GAP (PID2022-139308OA-I00) funded by MICIU/AEI/10.13039/501100011033/ and by ERDF, EU; Grant PRE2021-097001 funded by MICIU/AEI/10.13039/501100011033 and by ESF+ (predoctoral training grant associated to project PID2020-113230RB-C21); and Centro de Investigación de Galicia "CITIC", funded by the Xunta de Galicia through the collaboration agreement between the Consellería de Cultura, Educación, Formación Profesional e Universidades and the Galician universities for the reinforcement of the research centres of the Galician University System (CIGUS).

Author contribution

Conceptualization: AMO, CGR, DV; Data curation: AMO; Investigation: AMO, CGR, DV; Visualization: AMO; Software: AMO; Methodology: AMO, CGR, DV; Project Administration: CGR, DV; Software: AMO; Validation: AMO, CGR, DV; Experiments: AMO; Formal analysis: AMO, CGR, DV; Writing - original draft: AMO, CGR, DV; Writing - Review & Editing: AMO, CGR, DV; Funding Adquisition; CGR, CV

Conflict of interest

The authors have no competing interest to declare that are relevant to the content of this paper.

References

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, É., Hesslow, D., Launay, J., Malartic, Q., et al.: The falcon series of open language models. arXiv preprint arXiv:2311.16867 (2023)

Alemany-Puig, L., Esteban, J., Ferrer-i-Cancho, R.: The Linear Arrangement Library. A new tool for research on syntactic dependency structures. In: Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021), pp. 1–16. Association for Computational Linguistics, Sofia, Bulgaria (2021). https://aclanthology.org/2021.quasy-1.1

Abid, A., Farooqi, M., Zou, J.: Persistent anti-muslim bias in large language models. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 298–306 (2021)

Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623 (2021)

Berzak, Y., Huang, Y., Barbu, A., Korhonen, A., Katz, B.: Anchoring and agreement in syntactic annotations. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 2215–2224. Association for Computational Linguistics, Austin, Texas (2016). https://doi.org/10.18653/v1/D16-1239 . https://aclanthology.org/D16-1239

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901. Curran Associates, Inc., ??? (2020). https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Cai, Z.G., Haslett, D.A., Duan, X., Wang, S., Pickering, M.J.: Does chatgpt resemble humans in language use? arXiv preprint arXiv:2303.08014 (2023)

Chong, D., Hong, J., Manning, C.: Detecting label errors by using pre-trained language models. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 9074–9091. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (2022). https://aclanthology.org/2022.emnlp-main.618

Crothers, E., Japkowicz, N., Viktor, H.L.: Machine-generated text: A comprehensive survey of threat models and detection methods. IEEE Access (2023)

Crothers, E., Japkowicz, N., Viktor, H., Branco, P.: Adversarial robustness of neural-statistical features in detection of generative transformers. In: 2022 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2022). IEEE

Chiang, C.-H., Lee, H.-y.: Can large language models be an alternative to human evaluations? In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 15607–15631. Association for Computational Linguistics, Toronto, Canada (2023). https://aclanthology.org/2023.acl-long.870

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., *et al.*: Palm: Scaling language modeling with pathways. Journal of Machine Learning Research **24**(240), 1–113 (2023)

Dugan, L., Ippolito, D., Kirubarajan, A., Callison-Burch, C.: RoFT: A tool for evaluating human detection of machine-generated text. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 189–196. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-demos.25 . https://aclanthology.org/2020.emnlp-demos.25

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., Gardner, M.: Documenting large webtext corpora: A case study on the colossal clean crawled corpus. arXiv preprint arXiv:2104.08758 (2021)

Eldan, R., Li, Y.: Tinystories: How small can language models be and still speak coherent english? arXiv preprint arXiv:2305.07759 (2023)

Ferrer-i-Cancho, R.: Euclidean distance between syntactically linked words. Physical Review E **70**, 056135 (2004) https://doi.org/10.1103/PhysRevE.70.056135

Ferrer-i-Cancho, R., Gómez-Rodríguez, C., Esteban, J.L., Alemany-Puig, L.: Optimality of syntactic dependency distances. Physical Review E **105**(1), 014308 (2022)

Ferrer-i-Cancho, R., Liu, H.: The risks of mixing dependency lengths from sequences of different length. Glottotheory **5**(2), 143–155 (2014) https://doi.org/10.1515/glot-2014-0014

Futrell, R., Mahowald, K., Gibson, E.: Large-scale evidence of dependency length minimization in 37 languages. Proceedings of the National Academy of Sciences **112**(33), 10336–10341 (2015) https://doi.org/10.1073/pnas.1502134112

Fröhling, L., Zubiaga, A.: Feature-based detection of automated language models: tackling gpt-2, gpt-3 and grover. PeerJ Computer Science **7**, 443 (2021)

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al.: The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027 (2020)

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C.C.T., Del Giorno, A., Gopi, S., Javaheripi, M., Kauffmann, P., Rosa, G., Saarikivi, O., et al.: Textbooks are all you need. arXiv preprint arXiv:2306.11644 (2023)

Hartmann, J.: Emotion English DistilRoBERTa-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/ (2022)

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.d.L., Hendricks, L.A., Welbl, J., Clark, A., et al.: Training compute-optimal large language models. arXiv preprint arXiv:2203.15556 (2022)

Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., Kamar, E.: ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 3309–3326. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.acl-long.234 . https://aclanthology.org/2022.acl-long.234

He, X., Nassar, I., Kiros, J., Haffari, G., Norouzi, M.: Generate, annotate, and learn: NLP with synthetic text. Transactions of the Association for Computational Linguistics **10**, 826–842 (2022) https://doi.org/10.1162/tacl_a_00492

Ippolito, D., Duckworth, D., Callison-Burch, C., Eck, D.: Automatic detection of generated text is easiest when humans are fooled. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1808–1822. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.164 . https://aclanthology.org/2020.acl-main.164

Javaheripi, M., Bubeck, S., Abdin, M., Aneja, J., Bubeck, S., Mendes, C.C.T., Chen, W., Del Giorno, A., Eldan, R., Gopi, S., et al.: Phi-2: The surprising power of small language models. Microsoft Research Blog (2023)

Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)

Kumar, S., Balachandran, V., Njoo, L., Anastasopoulos, A., Tsvetkov, Y.: Language generation models can cause harm: So what can we do about it? an actionable survey. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pp. 3299–3321. Association for Computational Linguistics, Dubrovnik, Croatia (2023). https://aclanthology.org/2023.eacl-main.241

Kumar, V., Choudhary, A., Cho, E.: Data augmentation using pre-trained transformer models. In: Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems, pp. 18–26. Association for Computational Linguistics, Suzhou, China (2020). https://aclanthology.org/2020.lifelongnlp-1.3

Lucy, L., Bamman, D.: Gender and representation bias in GPT-3 generated stories. In: Proceedings of the Third Workshop on Narrative Understanding, pp. 48–55. Association for Computational Linguistics, Virtual (2021). https://doi.org/10.18653/v1/2021.nuse-1.5 . https://aclanthology.org/2021.nuse-1.5

Li, Y., Bubeck, S., Eldan, R., Del Giorno, A., Gunasekar, S., Lee, Y.T.: Textbooks are all you need ii: phi-1.5 technical report. arXiv preprint arXiv:2309.05463 (2023)

Liu, B., Bubeck, S., Eldan, R., Kulkarni, J., Li, Y., Nguyen, A., Ward, R., Zhang, Y.: Tinygsm: achieving¿ 80% on gsm8k with small language models. arXiv preprint arXiv:2312.09241 (2023)

Li, K., Hopkins, A.K., Bau, D., Viégas, F., Pfister, H., Wattenberg, M.: Emergent world representations: Exploring a sequence model trained on a synthetic task. arXiv preprint arXiv:2210.13382 (2022)

Liao, W., Liu, Z., Dai, H., Xu, S., Wu, Z., Zhang, Y., Huang, X., Zhu, D., Cai, H., Liu, T., et al.: Differentiate chatgpt-generated and human-written medical texts. arXiv preprint arXiv:2304.11567 (2023)

Landgrebe, J., Smith, B.: Making ai meaningful again. Synthese **198**, 2061–2081 (2021)

Liang, P.P., Wu, C., Morency, L.-P., Salakhutdinov, R.: Towards understanding and mitigating social biases in language models. In: International Conference on Machine Learning, pp. 6565–6576 (2021). PMLR

Liu, N.F., Zhang, T., Liang, P.: Evaluating verifiability in generative search engines. arXiv preprint arXiv:2304.09848 (2023)

Li, Z., Zhu, H., Lu, Z., Yin, M.: Synthetic data generation with large language models for text classification: Potential and limitations. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 10443–10461. Association for Computational Linguistics, Singapore (2023). https://doi.org/10.18653/v1/2023.emnlp-main.647 . https://aclanthology.org/2023.emnlp-main.647

Manning, C.D.: Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In: International Conference on Intelligent Text Processing and Computational Linguistics, pp. 171–189 (2011). Springer

Munir, S., Batool, B., Shafiq, Z., Srinivasan, P., Zaffar, F.: Through the looking glass: Learning to attribute synthetic text generated by language models. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 1811–1822. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.eacl-main.155 . https://aclanthology.org/2021.eacl-main.155

Martínez, G., Conde, J., Reviriego, P., Merino-Gómez, E., Hernández, J.A., Lombardi, F.: How many words does chatgpt know? the answer is chatwords. arXiv preprint arXiv:2309.16777 (2023)

Martínez, G., Hernández, J.A., Conde, J., Reviriego, P., Merino, E.: Beware of words: Evaluating the lexical richness of conversational large language models. arXiv preprint arXiv:2402.15518 (2024)

Nguyen-Son, H.-Q., Tieu, N.-D.T., Nguyen, H.H., Yamagishi, J., Zen, I.E.: Identifying computer-generated text using statistical analysis. In: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1504–1511 (2017). IEEE

OpenAI: GPT-4 Technical Report (2023)

Petrov, S., Das, D., McDonald, R.: A universal part-of-speech tagset. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pp. 2089–2096. European Language Resources Association (ELRA), Istanbul, Turkey (2012). http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf

Peng, B., Li, C., He, P., Galley, M., Gao, J.: Instruction tuning with gpt-4. arXiv preprint arXiv:2304.03277 (2023)

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., Launay, J.: The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. arXiv preprint arXiv:2306.01116 (2023)

Pillutla, K., Swayamdipta, S., Zellers, R., Thickstun, J., Welleck, S., Choi, Y., Harchaoui, Z.: Mauve: Measuring the gap between neural text and human text using divergence frontiers. Advances in Neural Information Processing Systems **34**, 4816–4828 (2021)

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A python natural language processing toolkit for many human languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 101–108. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-demos.14 . https://aclanthology.org/2020.acl-demos.14

Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, ??? (2019). https://arxiv.org/abs/1908.10084

Rodriguez, J., Hay, T., Gros, D., Shamsi, Z., Srinivasan, R.: Cross-domain detection of GPT-2-generated technical text. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1213–1233. Association for Computational Linguistics, Seattle, United States (2022). https://doi.org/10.18653/v1/2022.naacl-main.88 . https://aclanthology.org/2022.naacl-main.88

Rosenfeld, A., Lazebnik, T.: Whose llm is it anyway? linguistic comparison and llm attribution for gpt-3.5, gpt-4 and bard. arXiv preprint arXiv:2402.14533 (2024)

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020)

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., Hashimoto, T.: Whose opinions do language models reflect? arXiv preprint arXiv:2303.17548 (2023)

Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., Gallé, M., et al.: Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100 (2022)

Stribling, J., Krohn, M., Aguayo, D.: Scigen-an automatic cs paper generator (2005)

Sadasivan, V.S., Kumar, A., Balasubramanian, S., Wang, W., Feizi, S.: Can ai-generated text be reliably detected? arXiv preprint arXiv:2303.11156 (2023)

Søgaard, A.: Understanding models understanding language. Synthese **200**(6), 443 (2022)

Sahu, G., Rodriguez, P., Laradji, I., Atighehchian, P., Vazquez, D., Bahdanau, D.: Data augmentation for intent classification with off-the-shelf large language models. In: Proceedings of the 4th Workshop on NLP for Conversational AI, pp. 47–57. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.nlp4convai-1.5 . https://aclanthology.org/2022.nlp4convai-1.5

Swayamdipta, S., Schwartz, R., Lourie, N., Wang, Y., Hajishirzi, H., Smith, N.A., Choi, Y.: Dataset cartography: Mapping and diagnosing datasets with training dynamics. In: Proceedings of the 2020 Conference on Empirical Methods in Natural

Language Processing (EMNLP), pp. 9275–9293. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-main.746 . https://aclanthology.org/2020.emnlp-main.746

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html **3**(6), 7 (2023)

Tang, R., Han, X., Jiang, X., Hu, X.: Does synthetic data generation of llms help clinical text mining? arXiv preprint arXiv:2303.04360 (2023)

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open Foundation and Fine-Tuned Chat Models (2023)

Wei, J., Huang, D., Lu, Y., Zhou, D., Le, Q.V.: Simple synthetic data reduces sycophancy in large language models. arXiv preprint arXiv:2308.03958 (2023)

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al.: Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359 (2021)

Xu, Q., Peng, Y., Wu, M., Xiao, F., Chodorow, M., Li, P.: Does conceptual representation require embodiment? insights from large language models. arXiv preprint arXiv:2305.19103 (2023)

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems **32** (2019)

Zhan, H., He, X., Xu, Q., Wu, Y., Stenetorp, P.: G3detector: General gpt-generated text detector. arXiv preprint arXiv:2305.12680 (2023)

Zhou, J., Zhang, Y., Luo, Q., Parker, A.G., De Choudhury, M.: Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1–20 (2023)