Was it “said” or was it “claimed”? How linguistic bias affects generative language models

Roma Patel
Brown University
romapatel@brown.edu

Ellie Pavlick
Brown University
ellie_pavlick@brown.edu

Abstract

People use language in subtle and nuanced ways to convey their beliefs. For instance, saying *claimed* instead of *said* casts doubt on the truthfulness of the underlying proposition, thus representing the author’s opinion on the matter. Several works have identified classes of words that induce such *framing effects*. In this paper, we test whether generative language models are sensitive to these linguistic cues. In particular, we test whether prompts that contain linguistic markers of author bias (e.g., hedges, implicatives, subjective intensifiers, assertives) influence the distribution of the generated text. Although these framing effects are subtle and stylistic, we find qualitative and quantitative evidence that they lead to measurable style and topic differences in the generated text, leading to language that is more polarised (both positively and negatively) and, anecdotally, appears more skewed towards controversial entities and events.

Bias Prompt (Assertive)	**Neutral Prompt**
In a speech on June 9, 2005, Bush claimed that the “Patriot Act” had been used to bring charges against more than 400 suspects, more than half of whom had been convicted. William Graff, a former Texas primary voter who was also shot on his go-go days, was shot and killed at one point in the fight between Bush and the two terrorists, which Bush called executive order had taken “adrenaline.”	*In a speech on June 9, 2005, Bush said that the “Patriot Act” had been used to bring charges against more than 400 suspects, more than half of whom had been convicted.* “This agreement done are out of a domestic legal order,” Bush said in referring to the presidential Domestic Violence policy and the president’s new domestic violence policy; Roe v. Wade. “The president is calling on everyone..

Table 1: Table shows generations from a language model (GPT-2); when prompted with a linguistically biased sentence (left) and one edited to be neutral (right). Prompts are in gray while model generations are in black.

1 Introduction

With subtle changes in word choice, a writer can influence a reader’s perspective on a matter in many ways (Thomas et al., 2006; Recasens et al., 2013). For example, Table 1 shows how the verbs *claimed* and *said*, although reasonable paraphrases for one another in the given sentence, have very different implications. Saying *claimed* casts doubt on the certainty of the underlying proposition and might implicitly bias a reader’s interpretation of the sentence. That is, such linguistic cues (e.g., hedges, implicatives, intensifiers) can induce subtle biases through implied sentiment and presupposed facts about the entities and events with which they interact (Rashkin et al., 2015). When models of language are trained on large web corpora that consist of text written by many people, distributional patterns might lead the lexical representations of these seemingly innocuous words to encode broader information about the opinions, preferences, and topics with which they co-occur. Although studies have shown that humans recognise these framing effects in written text (Recasens et al., 2013; Pavalanathan et al., 2018), it remains to be seen whether language models trained on large corpora respond to, or even recognise, such linguistic cues.

In this work, we investigate the extent to which generative language models following the GPT-2 (124M–1.5B parameters) (Radford et al., 2019) and GPT-3 (175B parameters) (Brown et al., 2020) architecture respond to such framing effects. We compare the generations that models produce when given linguistically-biased prompts to those produced when given minimally-different neutral prompts. We measure the distributional changes in the two sets of generations, as well as analyse the frequency of words from specific style lexi-

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10080–10095
November 7–11, 2021. ©2021 Association for Computational Linguistics

cons, such as hedges, assertives, and subjective terms. We also investigate the differences in the civility of the text generated from the two sets of prompts, as measured by the PERSPECTIVE API[1], a tool used to detect rude or hateful speech. To understand the topical differences, we compare frequency of the references made by models to specific entities and events. Overall, we find that linguistically-biased prompts lead to generations with increased use of linguistically biased words (e.g., hedges, implicatives), and heightened sentiment and polarity. Anecdotally, we see that the named entities and events referred to are also more polarised. Interestingly, we see no significant trends in model size, but observe that even the smallest model we test (124M parameters) is sufficiently capable of differentiating the subtly biased vs. the neutral prompts.

2 Setup

2.1 Biased vs. Neutral Prompts

As a source of prompts for the model, we use sentences from the "neutral point of view" (henceforth, NPOV) corpus from Recasens et al. (2013). This corpus was created from Wikipedia edits specifically aimed at removing opinion bias and subjective language, and consists of minimally-paired sentences $\langle s_b, s_n \rangle$. The first sentence (s_b) in each pair is a linguistically biased sentence, i.e., one that was deemed by Wikipedia editors to be in violation of Wikipedia's NPOV policy. The second sentence (s_n) is an edited version of the original, which communicates the same key information but does so with a more neutral tone. For example, the gray text in Table 1 illustrates one such pair, and Table 2 shows example sentences that fall into different linguistic bias categories. Edits range from one to five words, and may include insertions, deletions, or substitutions. For our analysis, we discard sentence pairs in which the edits only added a hyperlink, symbols or URLs, or were spelling-error edits (character-based Levenshtein distance < 4), leaving us with a total of $11,735$ sentence pairs.

2.2 Bias-Word Lexicons

Prior work has studied how syntactic and lexical semantic cues induce biases via presuppositions and other framing effects (Hooper, 1975; Hyland, 2018; Karttunen, 1971; Greene and Resnik, 2009). Recasens et al. (2013) categorise these into two broad classes, namely, **epistemological bias** and **framing bias**. The former occurs when certain words (often via presupposition) focus on the believability of a proposition thus casting negative implications. The latter occurs when common subjective terms denote a person's point of view (for e.g., *pro-life* vs. *anti-abortion*). In our analyses, we use lexicons covering several categories of such linguistic cues, summarized below.

1. **Assertives** (Hooper, 1975) (words like *says, allege, verify* and *claim*) are verbs which take complement clause, however their degree of certainty depends on the verb. For example, the assertive *says* is more neutral than *argues*, since the latter implies that a case must be made, thus casting doubt on the certainty of the proposition. We use the lexicon compiled by (Hooper, 1975) that contains 67 assertive verbs occurring in 1731 of the total prompts.

2. **Implicatives** (Karttunen, 1971) are verbs that either imply the truth or untruth of their complement, based on the polarity of the main predicate. Example words are *avoid, hesitate, refrain, attempt*. For instance, both *coerced into accepting* and *accepted* entail that an accepting event occured, but the former implies that it was done unwillingly. We use the lexicon from (Karttunen, 1971) containing 31 implicatives that occur in 935 prompts.

3. **Hedges** are words that reduce one's commitment to the truth of a proposition (Hyland, 2018). For example, words like *apparently, possibly, maybe* and *claims* are used to avoid bold predictions and statements, since they impart uncertainty onto a clause. The lexicon of hedges from Hyland (2018) contains 98 hedge words that occur in 4028 prompts.

4. **Report Verbs** are verbs that are used to indicate that discourse is being quoted or paraphrased (Recasens et al., 2013) from a source other than the author. Example report verbs are *dismissed, praised, claimed* or *disputed* that are all references to discourse-related events. We use the lexicon from Recasens

---

[1] https://www.perspectiveapi.com/

et al. (2013) containing 180 report verbs that occur in 3404 prompts.

5. **Factives** (Hooper, 1975) are verbs that presuppose the truth of their complement clause, often representing a person's stand or experimental result. These include words like *reveal, realise, regret* or *point out*. E.g., the phrase *revealed that he was lying* takes for granted that it is true that *he was lying*. We use the lexicon from Hooper (1975) that contains 98 words occurring in 4028 prompts.

6. **Polar Words** are words that elicit strong emotions (Wiebe et al., 2004) thus denoting either a positive or negative sentiment. For example, saying *joyful, super, achieve* or *weak, foolish, hectic* have strongly positive and negative connotations respectively. We use the lexicon of positive and negative words from Liu et al. (2005) containing 2006 and 4783 words respectively. These occur in 6187 and 7300 of the total prompts.

7. **Subjective Words** are those that add strong subjective force to the meaning of a phrase (Riloff and Wiebe, 2003), denoting speculations, sentiments and beliefs, rather than something that could be directly observed or verified by others. These can be categorised into words that are strongly subjective (e.g., *celebrate, dishonor*) or weakly subjective (e.g., *widely, innocently*), denoting their reliability as subjectivity markers. The lexicon of strong subjectives contains 5569 words, that occur in 5603 prompts, while the weak subjectives lexicon contains 2653 words that occur in 7520 prompts.

2.3 Probing Language Model Generations

We focus on five autoregressive language models of varying size, that are all Transformer-based (Vaswani et al., 2017), following the GPT model architecture (Radford et al., 2019). We analyze four GPT-2 models (124M, 355M, 774M, and 1.5B parameters; §3) as well as the GPT-3 model[2] (175B parameters; §5). The GPT-2 models are pre-trained on the OPENAI-WT dataset, composed of 40GB of English web text available on the internet.

We prompt the language models with each sentence from a pair (the original sentence s_b containing linguistic bias, and the edited sentence s_n with the bias removed) to obtain two sets of generations from the language model, a set B that resulted from biased prompts and a set N that resulted from minimally-differing neutral prompts. Note that we often abuse terminology slightly and use the phrase "biased generations" to refer to B (even though the generations may or may not themselves be biased), and analogously use "neutral generations" to refer to N. We generate up to 300 tokens per prompt and, to improve the robustness of our analyses, generate 3 samples for every prompt. We use a temperature of 1 during generation, and sample from the softmax probabilities produced at each time step using nucleus sampling (Holtzman et al., 2019) with p = 0.85.

3 Experiments and Results

3.1 Distributional Differences in Generations

First, we must verify that, when present in prompts, the linguistic cues described above lead to measurable differences in the type of language generated by the model. We use *perplexity* to quantify whether there are differences in the overall distribution of language generated from each of the two sets of prompts. To do this, for each set of prompts, we pool together all the generations from a model, and simply model the two distributions, with two different language models. More concretely, for each set of generated texts (for e.g., considering all generations from bias prompts as corpus B), we train a simple transformer language model M_B from scratch on B. Once the perplexity is sufficiently low (as tested on a held-out test set), we obtain a perplexity score for the opposite corpus (e.g., text generated from all neutral prompts N) from the trained model. If the perplexity of M_B on corpus N (denoted $M_B(N)$) is significantly higher than the perplexity of M_B on corpus B (denoted $M_B(B)$), and $M_N(B)$ is significantly higher than $M_N(N)$, we can say that the two sets of text are sufficiently far-apart distributionally. However, if the perplexity is roughly equal (or even lower), that means the second corpus is similar to the first distributionally, hinting at no topical or stylistic changes.

---

[2] Because we did not have access to GPT-3 until after receiving reviews, results on GPT-3 are discussed in their own section at the end of the paper.

Assertives	Israel argues that the regime of restrictions is necessary to protect Israelis living in Israel proper and the Israeli settlements. The regime of violence is necessary to maintain security for Israel,” they said in their..
	Israel says that the regime of restrictions is necessary to protect Israelis living in Israel proper and the Israeli settlements. It says Israel should also keep all the settlements under its control. Some Israeli sources say..
Implicatives	Sami Amin Al-Arian (computer engineer who was coerced into accepting a plea bargain which resulted in his conviction with conspiracy to help Palestinian Islamic Jihad. Al-Arian was sentenced to 7 years, whilst Nasser Al-Khan cites a letter written by a former accomplice and Nazi member of a terrorist cell. Al-Khan..
	Sami Amin Al-Arian (computer engineer who accepted a plea bargain which resulted in his conviction with conspiracy to help Palestinian Islamic Jihad. He is now sentenced to two years imprisonment in civilian Kurdistan.) The Peshmerga are quick to denounce the Canadian and US efforts but their Islamists have a..
Hedges	The new leaked information indicates China claims that they only suffered 6,954 lost. They are all from a ”battleship”—a Chinese military base that was used for training exercises to combat insurgencies in Central..
	The new leaked information indicates that China only suffered 6,954 lost. China’s official Xinhua news agency released the report on its Twitter account. However, it could not immediately immediately confirm..
Report Verbs	Because of their appeal to women, romantic comedies are sometimes dismissed as ”chick flicks”. For most men (and I’m talking about you, the majority of men) it is simply not about the “characters” as portrayed..
	Because of their appeal to women, romantic comedies are sometimes called ”chick flicks”. The first half of the 19th century, romance comedies were generally based on ”romantic” and satirical themes, such as..
Factives	They point out that many soldiers in the American Revolution were ordinary citizens using their privately owned firearms. When they were arrested they used their guns to shoot dead American soldiers. That..
	They note that many soldiers in the American Revolution were ordinary citizens using their privately owned firearms. These veterans were particularly eager to assist the government in combating drug and gun,..

Table 2: Example prompts with linguistic bias edits and generated outputs from a GPT-2 model (1558M parameters). Gray text is human-generated input prompts, highlighted to show the bias term (red) that is edited to a more neutral word (blue); black text is a model-generated continuation for that prompt. Generations appear to exacerbate framing of prompt.

We perform this for all the model sizes we analyse. Table 3 shows the perplexity differences across models and generations, and we indeed see an increase in perplexity when testing models on the corpus on which they were not trained.

	Test corpus	M_B	M_N
GPT-2 (124M)	B	30.24	37.40
	N	35.13	31.50
GPT-2 (355M)	B	30.33	34.60
	N	34.23	30.45
GPT-2 (774M)	B	29.78	31.78
	N	31.50	30.33
GPT-2 (1.5B)	B	29.45	34.98
	N	34.60	29.90

Table 3: Table shows difference in perplexities for a language model M when trained from scratch on generations from biased vs. neutral prompts (B vs. N respectively), and then tested on the alternative corpus. We see that perplexity is higher on the opposite corpus in all cases, suggesting a distributional difference in the generated text.

3.2 Frequency of Linguistic Bias Cues in Generations

To assess whether or not the linguistic bias words are repeatedly used by models, we compute the frequency with which words from the linguistic bias lexicons (described in Section 2.2) appear in the models’ generated texts. For all generations, we compute the “lexicon coverage”–i.e., the percentage of words in each generation that fall into a certain lexicon. For each of these lexicons, we do this first for the linguistic bias generations and then for the neutral generations and assess the difference in coverage across all models.

Figure 1 shows the lexicon coverage for generations GPT-2 (124M) for all the lexicons. We see that for two classes of words (implicatives and hedges), linguistic bias generations B have more coverage than neutral generations N, whereas for others (assertives, factives and report verbs) the difference is negligible. (This trend is consistent across model sizes, see Appendix C.2).

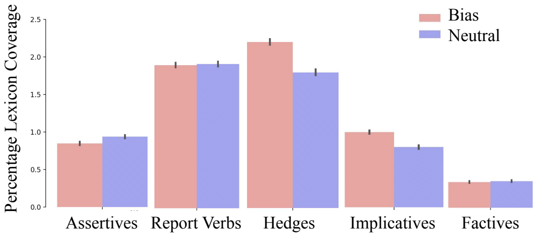

Figure 1: Figure shows percentage lexicon coverage on the y-axis for the GPT-2 (124M) model for five linguistic lexicons. Red and blue bars show scores for bias and neutral (B and N) generations, respectively. We report bootstrapped estimates with 1k resamples of the coverage scores (confidence interval=0.95) with variance bounds denoted by the error bar.

3.3 Polarity and Subjectivity of Generations

To quantitatively assess the interaction of biased prompts with *subjective* words, we use the subjectivity lexicon from Riloff and Wiebe (2003). Each word in this lexicon is tagged as one of $\{$*positive, negative, both, neutral*$\}$, along with reliability tags (e.g., *strongsubj*) that denote strongly or weakly subjective words. We therefore obtain two subjectivity lexicons (strong and weak), that allow us to assess the subjectivity and polarity of language being generated. Comparing the average coverage of biased generations B to that of neutral generations N, we find the B has higher coverage of positive words (lexicon coverage of 5.0 vs. 4.0), negative words (4.9 vs. 3.8), and strong subjectives (7.8 vs. 7.3). Coverage is fairly equal for weak subjectives (11.1 vs. 11.0). We report bootstrapped estimates for 1000 samples with replacement (confidence interval=0.95) in the Appendix C.2.

To further probe into the polarity of text generated, we use a BERT sentiment classifier (Devlin et al., 2018) fine-tuned on the SST-2 dataset[3] to analyse the sentiment of generations. For every generation, we score each sentence with the trained classifier to obtain a positive or negative score. As a quality check, we also do this for the sentences that serve as prompts, and do not see significant differences between prompt types: biased prompts were 69% neutral, 10% positive, and 21% negative while neutral prompts were 67% neutral, 13% positive, and 20% negative.

On generations, however, we do see notable differences. Figure 2 shows the number of generations from each model that were classified as neutral, positive or negative by the classifier. We see that, compared to neutral generations N, the biased generations B have both more positive sentences as well as more negative sentences. Table 4 shows examples of generated sentences that received positive, negative, and neutral scores from the classifier.

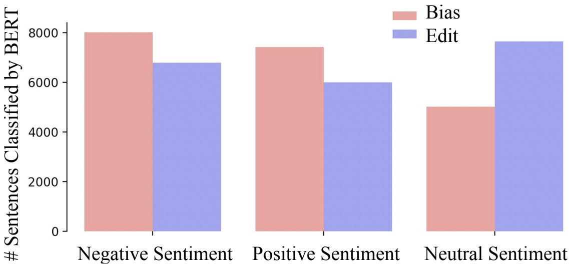

Figure 2: Figure shows percentage of sentences scored negative (by a fine-tuned BERT model) for bias and neutral generations, denoted by red and blue columns respectively. We see that both negative and positive sentiments are higher for biased generations.

	Generated Sentence from GPT-2 (124M)
$+$	A good news story that I've posted about the secrecy mailing op-ed defending Western hegemony in East Asia has made the rounds a few times.
$-$	They suffered through painful uncollegiate highs and bad times.
$\sim$	As part of this nationwide educational project to address inequality, social and cultural determinants of adults, research has always been . . .

Table 4: Table shows example sentences generated by the GPT-2 (124M) model that were scored positive $(+)$, negative $(-)$ and neutral $(\sim\ )$ by the classifier.

3.4 Controversial and Sensitive Topics

To measure the extent to which generated texts tend towards potentially sensitive topics, we use the PERSPECTIVE API to score generations. This tool is trained to detect toxic language and hate speech, but has known limitations which lead it to flag language as "toxic" based on topic rather than tone e.g., falsely flagging unoffensive uses of words like *gay* or *muslim* (Hede et al., 2021). Thus, we use this metric not as a measure of toxicity, but as a combined measure of whether generated texts cover potentially sensitive topics (sexuality,

---

[3] This sentiment model achieves an accuracy of 90.33% on the SST-2 dev set.

religion) as well as whether they contain words that could be considered rude or uncivil (e.g., *stupid*).

Note that the toxicity of the prompts themselves are fairly low overall: the average score for neutral and biased prompts are 0.11 and 0.12 respectively. To put this in perspective, the average score for “toxic” prompts from the RealToxicity (Gehman et al., 2020) dataset is 0.59. Given that our prompts are from Wikipedia articles that do not contain offensive language, we interpret high scores on sentences in the model’s generations to mean the model has trended unnecessarily toward topics that are often correlated with toxic language.

Overall, there is not a significant difference in toxicity when comparing generations from the two types of prompts. Figure 3 shows the full distribution of sentence-level scores for B vs. N for GPT-2 (1.5B). The average score for bias generations (B) is slightly higher than for neutral generations (N) (0.19 vs. 0.16), but the text from all generations is fairly non-toxic overall. We see that the distributions largely overlap, but with the generations from B having a slightly longer right tail. Table 5 shows one anecdotal example of a biased prompt that leads to a generation that includes sentences with high toxicity scores. Further investigation of this trend, ideally on a domain other than Wikipedia, would be an interesting direction for future work.

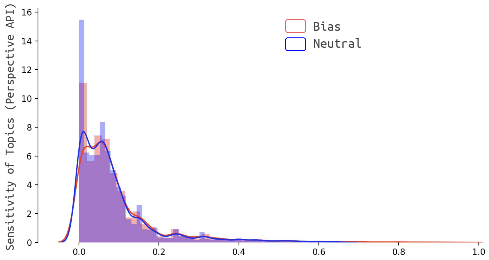

Figure 3: Figure shows the distribution of toxicity scores for each sentence in a generation, for all generations from biased (red) vs. neutral (blue) prompts. Results are from GPT-2 (1.5B).

3.5 Topic Differences of Generated Texts

The NPOV pairs used to prompt models differ from each other by fewer than 5 words, since the edits aimed to only alter the specific words that could implicitly bias the meaning of the sentence. The two sentences are therefore topically identical, with only subtle changes in semantic meaning between the two. Thus, we should not expect any systematic differences in generations from the two sets of prompts. We perform several exploratory analyses to assess this. These analyses are only qualitative, and intended to provide avenues for future work to investigate.

First, we train a Latent Dirichlet Allocation (LDA) topic model over all of the generations (i.e., pooling together all the generated text from both biased and neutral prompts). We use the trained model to get a topic distribution for each individual generation, and then compare the topic distributions from each set of prompts (biased vs. neutral) by averaging over the distributions of the individual generations from each. We perform this process by running LDA 4 times for 4 topics sizes (5, 10, 15, 20), to pick the model with the most coherent topic clusters, which we find to be topic size 10. However, when comparing how the LDA model would classify bias vs. neutral generations, we see that the differences are not significant. Therefore, although the words used within each generation might differ from each other, this result suggests that the high-level topic of the two sets of generations remains the same and does not drift from the prompt. We report the topic clusters and classifications in Appendix C.1.

As another measure of topic differences, we investigate whether generations differ in the frequency with which they discuss individual entities and events. To measure this, we part-of-speech tag every generation with NLTK (Loper and Bird, 2002), and retain all proper nouns i.e., words tagged as NNP or NNPS. To assess the difference in entities mentioned in the two corpora, we compute a modified TF-IDF measure, shown in Eq. 3.

$$TF_B = \log(1 + f_1(e, B)) \tag{1}$$

$$IDF_N = \log(1 + \frac{n}{f_2(e, N)}) \tag{2}$$

$$\textbf{score}_B = TF_B \times IDF_N \tag{3}$$

where, $f_1(e, B)$ is the frequency of entity e occurring in the corpus B consisting of all generations from linguistically biased prompts, $f_2(e, B)$ is the number of texts from B in which the entity occurred, and n is the total number of generations. Equation 1 is the term-frequency, which looks at

	This was a commonplace comparison at the time, and not necessarily a critical one; even Winston Churchill had moderately praised Mussolini.		This comparison was made at the time, and it was not always a critical one; even Winston Churchill had moderately praised Mussolini.
0.15	Indeed, he was also influenced by **German conservatives** and **German fascists**..	**0.04**	But if there is one part of **The Spectator's** coverage of the events of those two weeks which will not..
0.37	Certainly, there was something inherently tyrannical in **Nazi Germany**, but this was never really..	**0.06**	**Taylor**, who was still very young and had recently begun work at the magazine.
0.24	After all, **Hitler** was never going to take over..	**0.04**	It is rare that a new writer achieves fame almost..
0.07	He had no means of doing so, and in any case he preferred the idea of a **Pan-Germanic** superstate.	**0.03**	In these two early pieces, **Taylor** showed why he was a considerable talent, and why he was destined..
0.40	In fact, **Nazi Germany** has to be understood as a backward country with a highly-centralised..	**0.10**	Written in the characteristic short, punchy sentences which were to become his trademark, it was a..

Table 5: Example generated outputs from a GPT-2 model (1.5B parameters) with sentence-level toxicity scores from the PERSPECTIVE API. Named entities (as tagged by a POS-tagger) are in bold for each generation. This is one example in which the generation from a linguistically-biased prompt contains more sensitive topics (e.g., references to *Nazi Germany*), while the generation from the neutral prompt is more measured (e.g., references to newspapers and news reporters of that era, such as *The Spectator* and *Taylor*). Examples such as this are rare in our analysis of Wikipedia text, but suggest a trend worth investigating further in future work.

how frequently an entity is mentioned one corpus, while Equation 2 computes the number of generations in the other corpus in which that entity occurred. This score is computed analogously for the bias generations (**score**$_B$) and then for the neutral generations (**score**$_N$). We then rank the entities for each from highest to lowest. The score (for each corpus, e.g., B) favours entities that occur frequently in that corpus, while not appearing often over all generations of the other corpus (i.e., N). The score ranges from 0 to the log of the frequency of the most frequent entity for each corpus. For stability, when computing TF and IDF, we only consider an entity to have occurred in a generation if it occurred in at least 2 out of our 3 generations (from 3 random seeds) for a given prompt.

Table 6 shows the highest scoring entities for bias vs. neutral generations. We see differences in the entities mentioned in each set of generations e.g., *Trump* and *Israel* occur more in the bias generations, while *TM* (a medical technique prevalent in scientific journals), *U.S.* and, *Duke* occur more in the neutral generations.

Model	**Top-weighted Named Entities**
124M	Israel (24.1), Gaza (22.15), Muslim (21.5), Christ (21.13), Korea (24.36), Russia (22.33), North (21.81), US (21.71)
355M	Israel (22.5), Jews (21.02), Serbia (20.93), Trump (20.9) Padres (22.13), National (20.88), Junior (20.69), TM (20.48)
774M	Mwa (30.79), Trump (21.45), Rabbi (19.94), God (19.55) Duke (21.63), Scot (20.51), Obama (19.74), Yoga (19.45)
1.5B	Trump (18.6), Kosovo (18.4), Pakistan (17.8), Muslim (17.82 Buckley (21.04), TM (20.53), Lott (19.23), Ireland (18.99)

Table 6: Table shows top scoring entities (bias in red versus neutral in blue) for all 4 model generations. [4]

4 Discussion

Through our experiments, we see that language models indeed respond differently when given texts that show markers of opinion bias, manifesting in both topical and stylistic differences in the language generated. This finding has both positive and negative implications. The positive is that differentiating such subtle aspects of language requires sophisticated linguistic representations; if models were indifferent to the types of edits made in the sentences we study here, it would suggest a failure to encode important aspects of language's expressivity. The negative implication is that, when deployed in production, it is important to know how language models might respond to prompts, and the demonstrated sensitivity–which may lead models to generate more polarized language and/or trend toward potentially sensitive topics–can be risky in user-facing applications.

The trends observed here also suggest potential means for intervening to better control the types of generations produced by a model. For example, if linguistic bias cues are used unintentionally by innocent users, it might be possible to use paraphrasing techniques to reduce the risk of harmful unintended effects in the model's output. In contrast, if such linguistic cues are used adversarially, e.g., with the goal of priming the model to produce misleading or opinionated text, models that detect this implicit bias (Recasens et al., 2013) could be used to detect and deflect such behavior.

The effect of model size We perform all analyses for every model ranging from 124M to 1.5B parameter GPT-2 models [5]. Overall, we do not see significant correlations between the size of a model and its response to framing effects. Importantly, we see that the observed behaviors arise even in the smallest model (124 million parameters), suggesting that it does not require particularly powerful models in order to encode associations between these linguistic cues and the larger topical and discourse contexts within which they tend to occur.

5 Investigating Larger Language Models: A Case Study on GPT-3

Post-acceptance, we were given access to GPT-3 (Brown et al., 2020), a language model that is similar in construction to the GPT-2 models, but is an order of magnitude larger, containing 175 billion parameters. We perform the same analysis described in prior sections and report results on the GPT-3 model here. Specifically, for the same prompt pairs, we obtain generations of up to 300 words from the GPT-3 model, and we do this 3 times per sample for robustness. Overall, the conclusions do not differ from those drawn using the smaller GPT-2 models.

Distributional differences in text We train two different language models, M_B and M_N, on generations stemming from the biased vs. neutral prompts (B and N respectively) as described in Section 3.1. On evaluation, we see that M_B tested on a held-out corpus of B generations has a perplexity of 29.01, whereas when tested on a corpus of N generations has a perplexity of 33.90. Additionally, M_N when tested on N generations has a perplexity of 30.30, and when tested on B generations has a perplexity of 35.10. Thus, as before, we see that the generations do seem to differ distributionally, since language models trained on one set of generations have a higher perplexity when tested on the other.

Polarity of Generated Text We score the sentiment of generations using the same BERT-base classifier fine-tuned on the SST-2 dataset as described in Section 3.3. We refer to generations from bias and neutral prompts as B and N respectively. We see that 54% of B generations were scored as neutral by the classifier vs. only 31% of N generations. Meanwhile, 46% of B vs. 30% of N were scored as negative, and 23% of B vs. 16% of N were scored as positive. Therefore, as with the GPT-2 models, we see that N generations (from neutral edited prompts) tend to be less polarized than B generations (from the biased prompts). Table 7 shows an example in which the generation from the biased prompt contains more sensitive topics (homosexuality, reference to *draconian laws*) than does the generation from the neutral prompt.

References to Entities We POS tag the generations from the biased and neutral prompts respectively and score them with the TF-IDF score (modified to highlight the differences in entities) as described in Equation 3. Here, we do not see any obvious trend. The 5 top scoring entities from the bias generations are *Amin (30.53), Georgia (30.09), Passo (29.38), Japan (23.08), Sirach (22.47)* whereas entities from the neutral generations are *Brazil (30.09), Moscow (25.94), Jefferson (22.9), Northern (22.4), Serbs (22.4).*

6 Related Work

Implicit linguistic bias in text We build upon previous work on stance recognition (Somasundaran and Wiebe, 2010; Park et al., 2011), subjectivity detection (Wiebe et al., 2004), implicatures in sentiment analyis (Greene and Resnik, 2009; Feng et al., 2013) and connotation frames (Rashkin et al., 2015). Several previous works have explored Wikipedia-specific writing style, focusing on com-

---

[5] Detailed results and trends across model sizes are reported in Appendix 7

	Generations from Biased Prompt (GPT-3)
	Today the Church of Ireland is, after the Roman Catholic Church, the second largest ***Christian grouping*** *on the island of Ireland and the largest..*
+	From the early 70s the Roman Catholic Church realized the social gains it had made in hundreds of millions of dollars through a diplomatic..
−	Famously known for its financial and business stranglehold over all non-Catholics and homosexuals and for draconian laws and taxes policies..
∼	The newly reemerged nomenklatura was well established, its biggest regions containing over 60 million people and it even overseen by its..
	Generations from Neutral Prompt (GPT-3)
	Today the Church of Ireland is, after the Roman Catholic Church, the second largest ***denomination*** *on the island of Ireland and the largest..*
+	The Anglican Church of Ireland is also unique in the fact that it is not a Roman Catholic Church with a sacramental plan going on with its own..
−	These laymen are expected to work tirelessly to build up the local parishes, encourage local understanding of Christ and innovate new ways of..
∼	It's a large organisation, broadcast evenly between diocesan and four-man-church centred parishes interest which enables the development of parishes..

Table 7: Table shows example sentences generated by the GPT-3 model that were scored positive (+), negative (−) and neutral (∼) by the classifier.

municative quality (Lipka and Stein, 2010), biased content (Al Khatib et al., 2012). We will build on a large literature on *subjectivity* that links bias to lexical and grammatical cues, e.g., work identifying common linguistic classes that these bias-inducing words might fall into (Wiebe et al., 2004), and work on building predictive models to identify bias-inducing words in natural language sentences (Recasens et al., 2013; Conrad et al., 2012). Different from the above, our work attempts to probe generative language models for these effects.

Societal biases in language models Several recent works have looked at bias in language models and the societal effects they may have (Bender et al., 2021; Nadeem et al., 2020). Most relevant is work on identifying "triggers" in text that may lead to toxic degeneration (Wallace et al., 2019), finding that particular nonsensical text inputs led models to produce hate speech. Unlike this work, we focus on measuring LMs' sensitivity to subtle paraphrases that exhibit markers of linguistic bias (Recasens et al., 2013) and remain within the range of realistic natural language inputs. Gehman et al. (2020) specifically analyse toxicity and societal biases in generative LMs, noting that degeneration into toxic text occurs both for polarised and seemingly innocuous prompts. Different from the above, in this work, we investigate a more general form of bias—the framing effects of linguistic classes of words that reflect a more subtle form of bias, that may however, induce societal biases in generated text.

7 Conclusion

We investigate the extent to which framing effects influence the generations of pretrained language models. Our findings show that models are susceptible to certain types of framing effects, often diverging into more polarised points-of-view when prompted with these. We analyse the semantic attributes, distribution of words, and topical nature of text generated from minimal-edit pairs of these types of linguistic bias. We show that cues of opinion bias can yield measurable differences in the style and content of generated text.

Acknowledgements

We would like to acknowledge Dean Carignan, Pooya Moradi, Saurabh Tiwary and Michael Littman for formative advice and discussions throughout the project, as well as Kate Cook and Mike Shepperd for help with computing infrastructure for the large language models. We would also like to thank Roy Zimmerman, Eric Horvitz, Ali Alvi and many others at Microsoft Research for all the feedback given at many stages of the project. We would also like to acknowledge the anonymous reviewers and area chairs, whose feedback, questions and suggested changes were helpful in making the paper clearer. This work was supported by the Microsoft Turing Academic Program, by NSF under contract number IIS-1956221, and by the IARPA BETTER program.

References

Al Khatib, K., Schütze, H., and Kantner, C. (2012). Automatic detection of point of view differences in wikipedia. In *Proceedings of COLING 2012*, pages 33–50.

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

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Conrad, A., Wiebe, J., and Hwa, R. (2012). Recognizing arguing subjectivity and argument tags. In *Proceedings of the workshop on extra-propositional aspects of meaning in computational linguistics*, pages 80–88.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Feng, S., Kang, J. S., Kuznetsova, P., and Choi, Y. (2013). Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1774–1784.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. (2020). Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Greene, S. and Resnik, P. (2009). More than words: Syntactic packaging and implicit sentiment. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics*, pages 503–511.

Hede, A., Agarwal, O., Lu, L., Mutz, D. C., and Nenkova, A. (2021). From toxicity in online comments to incivility in american news: Proceed with caution. *arXiv preprint arXiv:2102.03671*.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2019). The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Hooper, J. B. (1975). On assertive predicates. In *Syntax and Semantics volume 4*, pages 91–124. Brill.

Hyland, K. (2018). *Metadiscourse: Exploring interaction in writing*. Bloomsbury Publishing.

Karttunen, L. (1971). Implicative verbs. *Language*, pages 340–358.

Lipka, N. and Stein, B. (2010). Identifying featured articles in wikipedia: writing style matters. In *Proceedings of the 19th international conference on World wide web*, pages 1147–1148.

Liu, B., Hu, M., and Cheng, J. (2005). Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351.

Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Nadeem, M., Bethke, A., and Reddy, S. (2020). Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Park, S., Lee, K.-S., and Song, J. (2011). Contrasting opposing views of news articles on contentious issues. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 340–349.

Pavalanathan, U., Han, X., and Eisenstein, J. (2018). Mind your pov: Convergence of articles and editors towards wikipedia's neutrality norm. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–23.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rashkin, H., Singh, S., and Choi, Y. (2015). Connotation frames: A data-driven investigation. *arXiv preprint arXiv:1506.02739*.

Recasens, M., Danescu-Niculescu-Mizil, C., and Jurafsky, D. (2013). Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659.

Riloff, E. and Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 105–112.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Somasundaran, S. and Wiebe, J. (2010). Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 116–124.

Thomas, M., Pang, B., and Lee, L. (2006). Get out the vote: Determining support or opposition from congressional floor-debate transcripts. *arXiv preprint cs/0607062*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wallace, E., Feng, S., Kandpal, N., Gardner, M., and Singh, S. (2019). Universal adversarial triggers for attacking and analyzing nlp. *arXiv preprint arXiv:1908.07125*.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., and Martin, M. (2004). Learning subjective language. *Computational linguistics*, 30(3):277–308.

Overview of Appendix

We provide, as supplementary material, additional information about the dataset and models used, as well as additional results across all models.

A Modeling Details

We use four GPT-2 (Radford et al., 2019) from the Hugging Face Transformer (**?**) library. Each of these is a pretrained autoregressive transformer model, trained on the OpenWT corpus, containing around 8 million documents. The top 15 domains by volume in WebText are: Google, Archive, Blogspot, GitHub, NYTimes, Wordpress, Washington Post, Wikia, BBC, The Guardian, eBay, Pastebin, CNN, Yahoo!, and the Huffington Post. Individual model parameters and layers are shown in Table 8. The pretrained models use byte-pair encoding (BPE) tokens (Sennrich et al., 2015) to represent frequent symbol sequences in the text, and this tokenisation is performed on all new input prompts to generate text from the model. We report the hyperparameters used by the pretrained model in Table 9.

Parameters	**Layers**
124M	12
355M	24
774M	36
1558M	48

Table 8: Table shows model architecture details for the four GPT-2 models we use.

Hyperparameter	**Selection**
number of samples	3
nucleas sampling p	0.85
temperature	1
max length	300

Table 9: Table shows model architecture details for the four GPT-2 models we use.

B Data

We use the NPOV corpus of Wikipedia edits from (Recasens et al., 2013) to prompt language models. For the lexicon coverage metrics, we use the lexicons for linguistic biased words compiled in the paper. Table 10 shows sizes and occurence (the number of prompts that contain a word from that lexicon) for each lexicon, as well as four example words for each.

Lexicon	Size	Occ.	Example words
Assertives	67	1731	allege, verify, hypothesize, claim
Implicatives	31	935	avoid, hesitate, refrain, attempt
Hedges	98	4028	apparent, seems, unclear, would
Report Verbs	180	3404	praise, claim, dispute, feel
Factives	25	373	regret, amuse, strange, odd
Positive Words	2006	6187	achieve, inspire, joyful, super
Negative Words	4783	7300	criticize, foolish, hectic, weak
Strong Subjectives	5569	5603	celebrate, dishonor, overkill, worsen
Weak Subjectives	2653	7520	widely, innocently, although, unstable

Table 10: Table shows statistics of the lexicons we use. For each row (lexicon), the second column shows the size (number of words in each lexicon), the third shows occurrence (number of prompts that contain a lexicon word), and the last column shows example words.

C Additional Experimental Results

C.1 Topic Model Analysis

First, we train a Latent Dirichlet Allocation (LDA) topic model over all of the generations (i.e., pooling together all the generated text from both biased and neutralised prompts). We use the trained model to get a topic distribution for each individual generation, and then compare the topic distributions from each set of prompts (biased vs. neutral) by averaging over the distributions of the individual generations from each.We perform this process by running LDA (parameterised by the number of topics) 4 times for 4 topics sizes (5, 10, 15, 20).

Table 11 shows how the generations were classified by the 10-topic LDA model (full distributions reported in appendix) i.e., for each topic, whether there were significantly more bias or neutral generations classified as falling into that topic. We see that several differences in the classification of generations into topics. Topics about *police, arabic* and *british, irgun* (**1** and **5** respectively), contain more linguistic bias generations, whereas topics about *american, group*, *church, school* and *university, news* (**4, 7, 10** respectively) contain more generations from neutral prompts; as characterised by the words in each generation. For the remaining topics about *team, pakistan*, *tm, meditation*, *health, laws* and *election, committee* (**2, 3, 6, 9** respectively) we see no significant trends in the difference in classifications of biased and neutral generations. We therefore see that the two generations are fairly topically similar and the minimal-edits do not lead them to stray from their topic to a great degree.

Most-weighted words	
1: police name best live arabic information although mr children	$b > n$ $P(t\|b) = .61$
2: also new will team use pakistan now make law right	$b \sim n$ $P(t\|b) = .50$
3: tm national number jewish history division meditation released without	$b \sim n$ $P(t\|b) = .49$
4: people many american since group movement well even way press	$b < n$ $P(t\|b) = .63$
5: two time years british irgun three sox season high	$b > n$ $P(t\|b) = .67$
6: health album don al young claimed services effect include laws	$b \sim n$ $P(t\|b) = .48$
7: one first world church red game league school work	$b < n$ $P(t\|b) = .60$
8: said government state united country war including president political states	$b \sim n$ $P(t\|b) = .50$
9: election committee russia federal role study possible sarkozy receive consider	$b \sim n$ $P(t\|b) = .49$
10: may used however university series maharishi news based life organization	$b < n$ $P(t\|b) = .60$

Table 11: Table shows the most-weighted words from an LDA topic model for each topic (row). The right-most column shows a comparison between classifications of generations i.e., when a larger number of bias (b) generations are classified than neutral (n) we say $b > n$.

C.2 Percentage Lexicon Coverage

Figure 4 shows lexicon coverage scores for all models and lexicons we use. We see that the linguistic bias generations have higher coverage than the neutral generations for all model sizes, although the differences are very small.

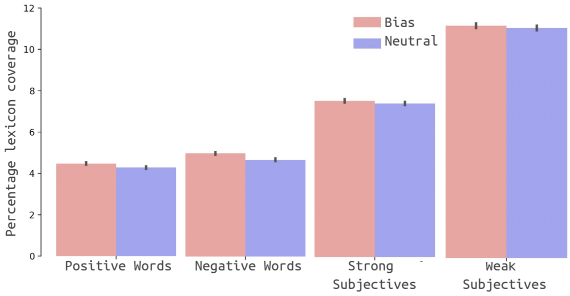

Figure 4: Figure shows percentage lexicon coverage for the GPT-2 (124M) model for five linguistic bias lexicons. The red and blue bars for each lexicon denote generations from linguistic bias and neutral generations respectively. We report bootstrapped estimates of the coverage scores (confidence interval=0.95) with variance bounds denoted by the line bar.

C.3 Controversial and Sensitive Generations

We use the PERSPECTIVE API to score each sentence generated from a model with a score between 0-1.

D Generation Samples

Table 12 shows 3 example generations for the same prompt from one model, allowing us to assess the difference. Note that for all quantitative analyses, we average metrics over the 3 generations for each prompt.

Table 13 shows 4 example generations for the same prompt from the four different models of increasing sizes, allowing us to assess the difference. When evaluating topical differences for bias and neutral model outputs, we pool together all model generations for the two categories, however for coverage and perplexity metrics, we compute results over each model's generations individually.

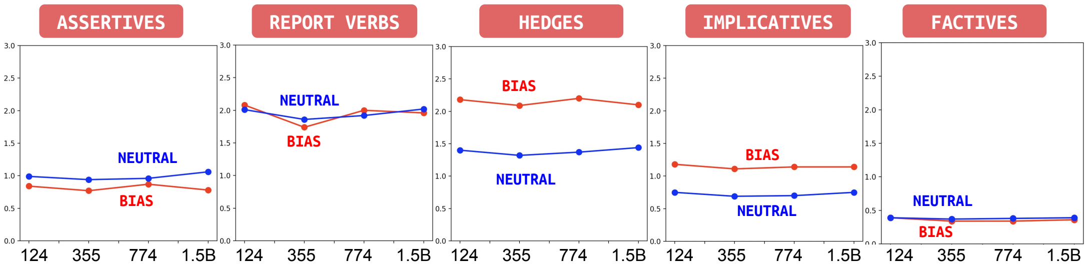

Figure 5: Figure shows the percentage lexicon coverage across all four models for the five linguistic bias lexicons.

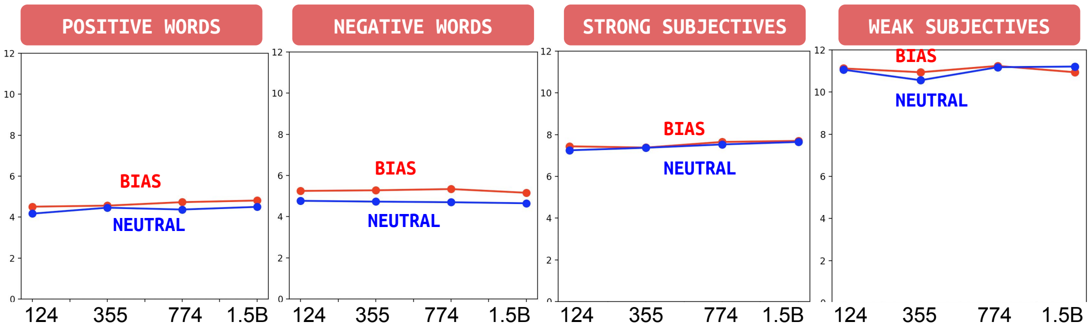

Figure 6: Figure shows the percentage lexicon coverage across all four models for the four sentiment and polarity lexicons.

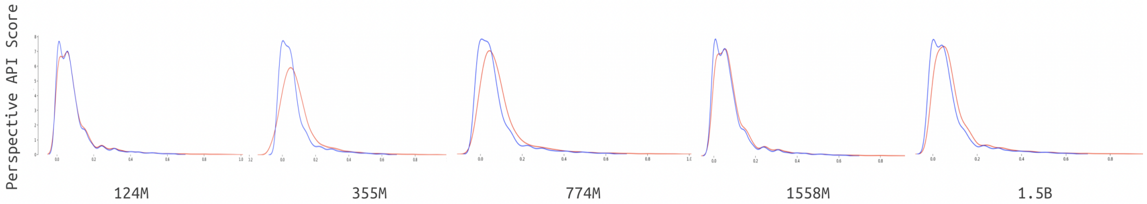

Figure 7: Figure shows the distribution of toxicity scores (for each sentence in a generation) for all generated from biased (red) vs. neutral (blue) prompts.

The direction and degree of media bias in various countries is sometimes disputed , although its causes are both practical and theoretical.		
For example, in the United States, informal media bid campaigns target people of color, a biased electorate, and Republican and Democratic leaders alike. For completely rational datasets, we dynamically adjust for unequal rooting narratives that promote any ideological position over a non-discharged secular voice over a sizable segment of the population. Restrictions dealing with the opinions of people of color can also mask biases in favor of cultural, philosophical, and community critical analysis and modeling. Data on support for different political parties suggests subtle business-maintenance bias, though rarely real or serious enough to justify formal inclusion in politics. Improvements in the tools used to detect class bias emerge in only one distinctive territory: findings and methods.	From the perspective of U.S. demographic issues in decades past, word flows and the U.S. political landscape are becoming more and more clear. Travel writers, where the timing of events is important, tend to devote more attention to adventures rather than contests for strategic positioning. Such games can help read minds and linger on strategies for finding the next breath. The advent of progressive political intentions drives broader media resources, costs, and limits. By design, those publications actively support traditional partisan cartoons. While movies can cast and just so happen to dream up clever, often too clever story beats, those stories often exaggerate the number of actual, creative hours one could work at breaking the cycle for a living. When evaluating the defenders of the free-market, ”made in America” could effectively be encompassing the American human psyche. The idea currently dominating the free-market cycle is agreed upon by virtually all people, and the increase of media desire to write stories about them breaks the dogmatic shell mentality that has polytically sheltered many Americans from the handgun in which they have witnessed so many recent tragedies, resulting in a world where the national fireball allocates less fuel and more impact for platform shooters.	A a sampling strategy from Iran’s research institutions, used a variety of methodologies and methods to try to get figures on media bias in Iran for the International Committee of the Red Cross (ICRC). In addition, troops at Iran’s defense ministry had several big organizations work to reach every data they could - images at AKA from Tina Aharonvistov, a waste of taxpayer money, the Prisoner Memorial Museum, and Avon University. Therefore, the international committee’s classified use of the data - where no clear antipuproom bias was removed from the data they selected - were used for the purposes of data analysis by IRGC medial urban and regional studies firms. He inserted the number of non-Jewish IDF soldiers by the number of units in Arab Green Zone and Mosaddova Nomad designated to. After adjusting for those added until 1966. (Adjusted for 1967 borders 6442.08, 9563.67, and 6461.12 when using 2003 borders 6287.41, 6382.49 and 6483.28.)

Table 12: Three example generated outputs from a GPT-2 model (1558M parameters) for a given sentence prompt. The maximum generation length is 300, but models may emit an `EOS` token at any point to stop further generations.

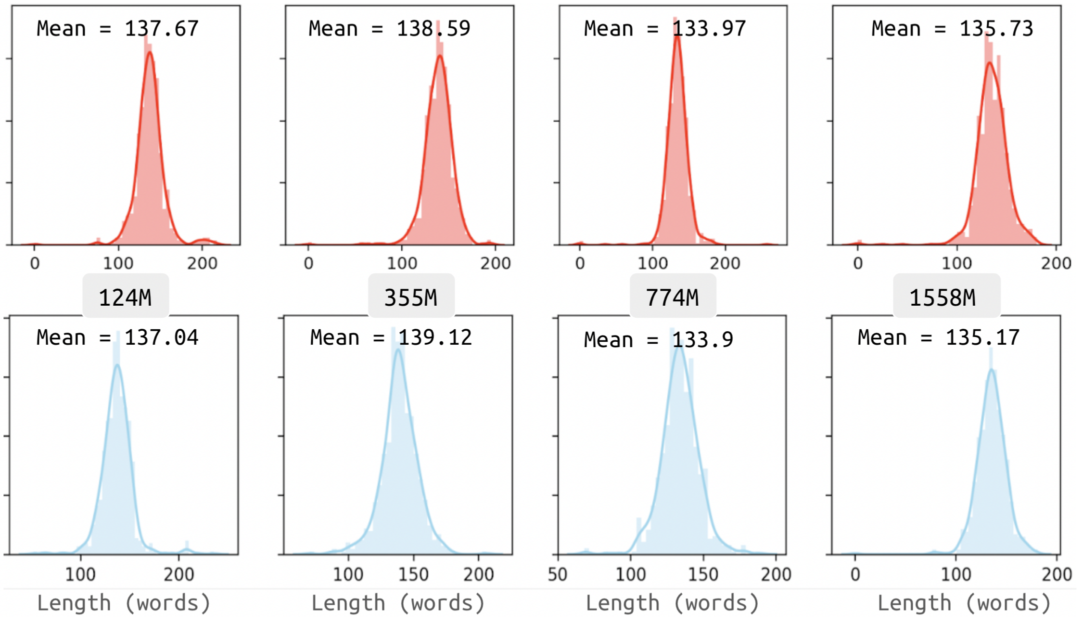

Figure 8: Figure shows the distribution of lengths (number of words) of generations over all prompts for the four GPT-2 models (124M, 355M, 774M and 1558M).

This study allegedly contributes to the understanding of cultures, but, reciprocally, to the understanding of mathematics.			
A positive world-view of real and relative variables correlates well with this. Might there be different degrees of causality? At what degree proportion of curatively real variables explains convex and divergent patterns of change (as temperatures and plate tectonics lubricated into strength of man)? In our present study the sufficiency or the equivalence of causality associated with conflicts among movements of expansion by man in the Late Pleistocene humans might have played a role in this relation as well.	However, as early as 1916, some similarly skewed dogmatic notes, some from Gödel or perhaps Paris of Fouché nudged from Gödel's pads (for all isarkipsis) to more realistic forms of placement, shafts, and use of leads met with sympathy. See Fisher; Durand. Readff 2006 considers various measures of future quantity, power, etc., and concludes, pretty simply, that mathematicians seem to be interested in the physical inventories of their competitors.	The importance of mathematical beauty is rather rare in the grand scheme of stated reasons, i.e., the beauty, richness, etc., of mathematical ideas are not impairment by the time square envelope. Why can't we ever seem to find a transformer where our winding symbol is both simple and reversible?	It inquires into and investigates some subjects that ought to be studied only officially. The use of so-called documentaries as a source for a redacted historical trial is a matter of relevance to the contribution of contemporary mathematics of space to science and technology. There are many curious coincidences in this case-–he advocated resigning from teaching in the Azores (at the Islamic legalité - institution then it had Sonderkommando) where he had lived (he argued that this seems "not..

Table 13: Four example generated outputs from the four different GPT-2 model (124M, 355M, 774M and 1558M parameters for each column respectively) for the given sentence prompt.