Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection

Suchin Gururangan[†] **Dallas Card**[◇] **Sarah K. Dreier**[♡] **Emily K. Gade**[♣]
Leroy Z. Wang[†] **Zeyu Wang**[†] **Luke Zettlemoyer**[†] **Noah A. Smith**[†♠]
[†]University of Washington [◇]University of Michigan [♡]University of New Mexico
[♣]Emory University [♠]Allen Institute for AI
{sg01,zwan4,lsz,nasmith}@cs.washington.edu dalc@umich.edu
skdreier@unm.edu emily.gade@emory.edu lryw@uw.edu

Abstract

Language models increasingly rely on massive web crawls for diverse text data. However, these sources are rife with undesirable content. As such, resources like Wikipedia, books, and news often serve as anchors for automatically selecting web text most suitable for language modeling, a process typically referred to as *quality filtering*. Using a new dataset of U.S. high school newspaper articles—written by students from across the country—we investigate whose language is preferred by the quality filter used for GPT-3. We find that newspapers from larger schools, located in wealthier, educated, and urban zones (ZIP codes) are more likely to be classified as high quality. We also show that this quality measurement is unaligned with other sensible metrics, such as factuality or literary acclaim. We argue that privileging any corpus as high quality entails a language ideology, and more care is needed to construct training corpora for language models, with better transparency and justification for the inclusion or exclusion of various texts.

1 Introduction

The language models central to modern NLP are trained on large Internet corpora, typically gathered from community resources (e.g., Wikipedia; Liu et al. 2019) or web crawls (e.g., WebText, Common Crawl; Radford et al. 2019, Brown et al. 2020). The selection of texts impacts every research or deployed NLP system that builds on these models. Yet there is rarely any clear justification given for why various texts were included.

Web dumps like Common Crawl offer the promise of more diverse text than what is available in curated resources. However, much of the web consists of frequently replicated boilerplate (e.g., privacy policies), code (e.g., HTML and Javascript), pornography, hate speech, and more. Automated approaches, typically referred to as **quality filters**, are often applied in an effort to remove this undesirable content from training data.[1] These filters include code removers (Gao et al., 2020), heuristics (Rae et al., 2021), stopwords (Raffel et al., 2020), and classifiers (Brown et al., 2020; Wenzek et al., 2020).

Although quality filtering is often treated as a relatively neutral preprocessing step, it necessarily implies a value judgment: which data is assumed to be of sufficiently high quality to be included in the training corpus? More concretely, when a quality filter is a classifier trained on instances assumed to be of high (and low) quality, the selection of those examples will impact the language model and any downstream technology that uses it. Many filters use Wikipedia, books, and newswire to represent high quality text. But what texts are excluded as a result? Because natural language varies with social and demographic variables (Rickford, 1985; Eckert, 1989; Labov, 2006; Blodgett et al., 2016; Hovy and Yang, 2021; Lucy and Bamman, 2021, *inter alia*), we can also ask *whose* language will be excluded.

We begin with a summary of the handful of data sources used to construct training corpora for many language models and assumed to be of high quality (§2). The systematic authorship biases in these datasets motivate the study that follows, in which we replicate the quality filter from Brown et al. (2020). We apply this filter to a new dataset of U.S. high school newspapers, augmented with demographic data from the U.S. Census and the National Center for Education Statistics (§3). We demonstrate that the filter has strong topical and stylistic preferences, and favors text from authors who originate from regions with better educational attainment, urban centers, larger schools, and higher valued homes.

In sociolinguistics, the term **language ideology**

[1] We note that the term *quality* is often ill-defined in the NLP literature. For example, Brown et al. (2020) and Wenzek et al. (2020) refer to "high-quality text" or "high-quality sources"—both citing Wikipedia as an example—but without explaining precisely what is meant.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2562 - 2580
December 7-11, 2022 ©2022 Association for Computational Linguistics

refers to common (but often unspoken) presuppositions, beliefs, or reflections about language that justify its social use and structure (Craft et al., 2020). Our analysis helps to characterize the language ideology encoded in the quality filter used by Brown et al. (2020), a representative of a wider set of filtering methods. We also observe in §4 that the filter is unaligned with other plausible notions of quality: factuality ratings for news sources, standardized test scores, and literary awards. Of course, these institutions entail their own language ideologies. We argue that when constructing a corpus, one cannot avoid adopting some language ideology; appropriate choices will depend on the goals of the work, and one language ideology may conflict with another. In short, there is no truly general-purpose corpus.

Our code and analysis are publicly available.[2]

2 Motivation: Data Sources

Across the many language models recently reported in the literature, the same small group of datasets have been routinely used as training corpora—Wikipedia, collections of books, and popular online articles (§A.1). These data are often treated as exemplars of high quality text (Devlin et al., 2019; Liu et al., 2019; Radford et al., 2019; Raffel et al., 2020; Brown et al., 2020). Although these datasets include text from many sources, extensive research suggests that the voices they represent are drawn from a relatively small, biased sample of the population, over-representing authors from hegemonic social positions.

Wikipedia Wikipedia serves as a backbone for language models because of its scale, ease of use, permissive license, and goal of providing comprehensive coverage of human knowledge. However, although anyone can edit Wikipedia content, not everyone does. In practice, there are significant biases in Wikipedia authorship, content, and perspectives. For instance, despite efforts by Wikimedia, the site has been unable to resolve a persistent gender imbalance among its editors (Huang, 2013; Meta-wiki, 2018). This imbalance is reflected in who gets written about, and how (Bamman and Smith, 2014; Graells-Garrido et al., 2015; Wagner et al., 2015). There is also a pervasive urban bias; editors are less likely to come from rural areas, and coverage of these areas in Wikipedia tends to be more limited (Mandiberg, 2020). Although coverage in English Wikipedia is not limited to those places where English is a majority language, an Anglo-American perspective dominates coverage.[3] Lastly, a relatively small number of people are responsible for most of the content (Panciera et al., 2009; Matei and Britt, 2017). Wikipedia is thus less representative of language of the population than one might expect given its size and design.

Books Language models are also frequently trained on book corpora. BERT (Devlin et al., 2019) used the Toronto BookCorpus (Zhu et al., 2015), which consists of 7,185 self-published novels, a dataset criticized for copyright violation, imbalanced representation, and lack of documentation (Bandy and Vincent, 2021).

GPT-3 (Brown et al., 2020) and The Pile (Gao et al., 2020) both use much larger corpora of books (although the former do not identify the source of this data). However, the Pile's books (also called Books3) are not a random selection. Rather, they appear to be drawn from a torrent file containing hundreds of thousands of copyrighted eBooks.

Books3 deserves a more thorough investigation, but preliminary analyses reveal that the most prevalent authors in the corpus are prolific American and British writers, especially of romance, mystery, and children's books (e.g., Danielle Steel). This pattern should be considered against the backdrop of the American book publishing industry, which has been widely criticized as homogeneous (Lee & Low Books, 2020.[4])

News and Other Popular Internet Content Radford et al. (2019) scrape text from the websites featured in popular Reddit submissions (i.e., those that received at least three upvotes) to construct the training data for GPT-2. As the original corpus is unavailable, we analyze its open-source replica, OpenWebText (Gokaslan and Cohen, 2019).

We do not expect the corpus to represent a wide range of language variation; Reddit users are mostly male, younger, and liberal-leaning, which influences the types of content shared and upvoted on the platform (Barthel et al., 2016). Indeed,

[2] http://github.com/kernelmachine/quality-filter

[3] For example, of the ten most frequently mentioned people in English Wikipedia, seven are U.S. Presidents, two are prominent figures in Christianity, and the only woman is the British monarch, Queen Victoria.

[4] This 2020 U.S. study found that Black people comprise only 5% of the industry, and books by men tend to generate disproportionately more attention than those by women.

we find that 1% of the 311K unique top-level domains in OpenWebText contribute 75% of documents in the corpus (§A.2). The most common websites in OpenWebText are internationally circulating British and American news outlets (e.g., *BBC*, *New York Times*), blogging platforms (e.g., *Tumblr*, *Blogspot*), sports content (e.g., *ESPN*, *SB-Nation*), and tech news (e.g., *TechCrunch*, *Wired*). As expected, these links tend to appear on the most highly trafficked subreddits (e.g., */r/politics*, */r/worldnews*, */r/news*).

These news sources are likely dominated by formal writing styles from a relatively homogeneous set of authors (Arana, 2018; Grieco, 2018). The adherence to slowly evolving style guides expresses specific linguistic standards (Froke et al., 2020) and even geopolitical interests (Vultee, 2012), which encourage rules on language use that can reinforce gender norms and racial hierarchies (DiNicola, 1994; Bien-Aimé, 2016). Researchers find a striking lack of diversity in newsrooms and newspaper leadership.[5] This may be compounded by the economic hardships aspiring journalists must incur,[6] which act as a filter for who can afford to be employed in the news industry.

Summary These descriptive findings suggest that a disproportionate amount of text in the core data sources of existing language models is written by authors from select, relatively powerful social positions. Such text sources appear to favor privileged segments of the English-speaking population, including men; white populations; communities of higher socio-economic status; and people harboring American and Western European historical, geopolitical, and cultural perspectives. The resulting corpora tend to be less inclusive of the voices of women and members of marginalized groups. A likely implication may be that alternative perspectives, including those of people from rural areas; non-dominant gender, sexual, or racial identities; and counter-hegemonic vantage points, may be less likely to be included, and thus less likely to influence models trained on this data.

Although formal, streamlined content like news or Wikipedia articles may seem like desirable sources for high quality content, not all writing styles or substantive topics that might be relevant to language technologies and their user communities are represented in the resulting corpora. When deployed, however, many of the technologies using language models trained on these mainstream data will face language that—despite being less formal, professional, or carefully edited—is no less high quality and is essential to the communicative lives of the people who use it.

3 Measuring the Language Ideology of the GPT-3 Quality Filter

Empirically evaluating the full distribution of authors in the data sources from §2 is difficult, due to their size and the lack of metadata about each document's authors. We instead curate a new dataset of U.S. high school newspaper articles that varies both topically and along demographic variables that can be resolved using ZIP codes. Although we do not directly consider individual authors of these articles, this dataset is useful, in that it can be associated with extensive metadata at the level of individual newspapers. We then analyze the behavior of a (replicated) quality filter on text from this dataset and discuss its implications.

3.1 U.S. School News

Background Many U.S. schools produce a newspaper to give students journalism experience, to report on local news, to comment on national or global events, and to publish school-related material (e.g., announcements, campus life, student interviews, sports or honor rolls; Gibson, 1961). Because a school's access to resources is shaped by local income levels (Betts et al., 2000) and tied to student achievement (Greenwald et al., 1996), we expect schools in wealthier areas (relative to poorer areas) to produce newspaper content that is more similar to the formal, professional texts that a quality filter is likely to classify as high quality.

Collection We collect articles from English-language U.S. school newspapers that used a common Wordpress template.[7] After identifying 2483 schools who use this template, we scrape 1.95M articles from their respective newspaper sites (more details in §A.3). We retrieve article categories by extracting them from the article URL slugs. We then match each school to its population zone (ZIP

[5] As of 2018, racial minorities make up 37% of the U.S. population but only 17% of staff and 13% of leadership in U.S. newsrooms (Arana, 2018).

[6] In 2020, median salary for U.S. news analysis, reporters, and journalists was $35,950, a slight decrease from 2012 after adjusting for inflation: `https://pewrsr.ch/3qCO75v`

[7] `SNOsites.com`

code) using the Google Maps Place API.[8] We restrict our dataset to articles from U.S. high schools. We only consider articles from 2010–2019, remove pages under the *video*, *photo*, or *multimedia* categories, and remove schools that have less than 100 articles (which tend to contain scraping errors). The final corpus includes 910K articles, from 1410 schools, located in 1329 ZIP codes (552 U.S. counties) dispersed across all U.S. states (and the District of Columbia).

3.2 The GPT-3 Quality Filter

To investigate how quality correlates with various attributes of a newspaper, we re-implement the Brown et al. (2020) quality filter based on the description provided in the paper. The filter is a binary logistic regression classifier using n-gram features, trained to distinguish between reference corpora (Books3, Wikipedia, and OpenWebText) and a random sample of Common Crawl.

We replicate the filter as closely as possible using `scikit-learn` (Pedregosa et al., 2011), which we release, along with a demo.[9] To create the training data for the classifier, we sample 80M whitespace-separated tokens of OpenWebText, Wikipedia, and Books3 each for the positive class, and 240M whitespace-separated tokens of a September 2019 Common Crawl snapshot for the negative class.[10] We perform a 100-trial random hyperparameter search, fixing only the hashing vectorizer and basic whitespace tokenization, following the implementation in Brown et al. (2020). Our final classifier gets 90.4% F_1 (91.7% accuracy) on a held-out test set (§A.4). We then apply the quality filter to the U.S. SCHOOL NEWS data, computing a quality score per document, which we denote $P(\text{high quality})$.

3.3 Document-Level Analysis

We first explore document-level preferences of the filter. The GPT-3 quality filter is more likely to classify high school newspaper articles as low quality, compared to general newswire (§A.5).[11] This is unsurprising, since the training data for the GPT-3 quality filter included texts by professional journalists. §A.6 shows a random sample of text from the dataset with high and low quality scores, illustrating differences in style and formality.

Dependent variable: $P(\text{high quality})$

Feature	**Coefficient**
Intercept	0.471^{***}
Topic 5 (*christmas, dress, holiday*)	-0.056^{***}
Topic 2 (*school, college, year*)	-0.037^{***}
Topic 6 (*student, school, class*)	-0.004
Topic 1 (*people, just, like*)	0.003
Topic 7 (*movie, film, movies*)	0.062^{***}
Topic 3 (*music, album, song*)	0.113^{***}
Topic 4 (*people, women, media*)	0.197^{***}
Topic 9 (*game, team, players*)	0.246^{***}
Topic 8 (*Trump, president, election*)	0.346^{***}
Presence of first/second person pronoun	-0.054^{***}
Presence of third person pronoun	0.024
$\log_2$(Number of tokens)	0.088^{***}
R^2	0.336
adj. R^2	0.336

Table 1: Regression of the quality score of an opinion piece in the U.S. SCHOOL NEWS dataset, on document features ($N = 10$k). We observe that political and sports-related topics, the lack of first and second person pronouns, and longer document lengths are associated with higher quality scores. We omit topic 0 (*food*, *restaurant*, *eat*) to avoid a saturated model. See §A.7 for quality scores per topic. $^{***}p < 0.001$.

More notably, controlling for article category (e.g., opinion pieces), we find that the GPT-3 quality filter has apparent topical and stylistic preferences. For topical features, we train a topic model (via latent Dirichlet allocation; Blei et al. 2003) over opinion pieces with 10 topics. We also consider whether documents contain first, second, or third person pronouns, and the length of the document. We then combine these features in a regression to assess the effect of certain attributes on the document quality score, while controlling for other attributes.

The results of our regression are displayed in Table 1. We find that certain topics have quite large effect sizes (see §A.7 for the distribution of quality scores per topic). For example, documents entirely about former U.S. President Trump and the 2016 presidential election have quality scores 35 percentage points higher, on average, than the omitted topic about food, whereas documents about sports are 25 percentage points higher, relative to the omitted topic. Stylistically, the presence of first or second pronouns in a document decreases quality score by 5 percentage points, while a doubling

[8] `https://developers.google.com/maps/documentation/places/web-service/search-find-place?hl=en`

[9] `https://huggingface.co/spaces/ssgrn/gpt3-quality-filter`

[10] We download the Common Crawl snapshot using code provided by Wenzek et al. (2020).

[11] Here, the general newswire are articles from popular online news sources; see §4 for data details.

of the number of tokens in a document increases the quality score by 9 percentage points.

3.4 Demographic Analysis

We also examine whether the GPT-3 quality filter prefers language from certain demographic groups over others. We first check raw correlations between average quality scores (per newspaper) and features of interest. As in §3.3, we then combine the features in a regression model.

Demographic Features As discussed in §3.1, we expect *a priori* that content from schools located in wealthier, more educated, and urban areas of the U.S. will tend to have higher quality scores, relative to poorer, less educated, rural areas. Therefore, we consider demographic features that correspond to class, rural/urban divides, and school resources.

For each school, we retrieve 2017–2018 school-level demographic data from the National Center for Education Statistics (NCES).[12] These include the number of students, student:teacher ratio, and indicators for private schools and specialized public schools (e.g., charter or magnet schools). We also retrieve the latest ZIP code- and county-level demographic data from the 2020 U.S. Census.[13] To measure the wealth of the corresponding ZIP code, we use median home values, and for educational attainment we use the percentage of college-educated adults. We also use Census data on the percent of rural population by county. Finally, we consider local political leanings, operationalized by county-level Republican-party vote share in the 2016 Presidential election.[14] We display full descriptions of features in our demographic analysis in §A.8.

Correlation Analysis To inform the variables we include in our regressions, we explore correlations between variables of interest and the average quality score of a school newspaper. Our analyses in Figure 1 suggest that our initial hypotheses hold: schools in wealthier, urban, and more educated ZIP codes, as well as those in Democrat-leaning counties, tend to have higher quality scores.

Regression Analysis Here, we use schools as the unit of analysis, and consider average quality score assigned to the school's articles as the dependent variable. We only include those schools that could be matched to the NCES database, dropping schools which are missing school size, as well as those located in ZIP codes with $1M or greater median home value, due to a census artifact.[15] Missing values for other features are imputed with the median value of that feature for the corresponding ZIP code, or (if necessary) county or state. For regressions, we log-transform school size, student:teacher ratio, and home values, using raw values for other features to preserve interpretability. Our regression dataset includes 968 high schools in 926 ZIP codes across 354 counties. We release this dataset publicly.[16]

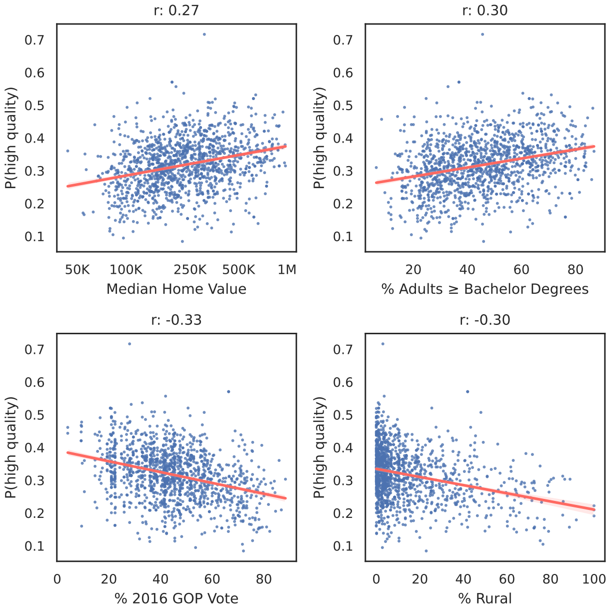

Figure 1: Scatter plots displaying correlations of select demographic features of a school's ZIP code or county with its average $P(\text{high quality})$.

Because many of the variables identified above are correlated, we use regression to estimate the effect of certain factors while controlling for others, with results shown in Table 2. Overall, home values, parental education, school size, public school status, and urban locations all show significant positive associations with quality scores. Thus, even controlling for financial resources, parental education, and other factors, articles from urban schools are scored as significantly higher quality than those from rural schools.

[12] https://nces.ed.gov/ccd/elsi/tablegenerator.aspx

[13] https://data.census.gov/cedsci/

[14] https://electionlab.mit.edu/data

[15] The census data has an artificial upper bound of $1M.

[16] https://github.com/kernelmachine/quality-filter

Dependent variable: P(high quality)

Feature	Coefficient
Intercept	0.076
% Rural	-0.069^{***}
% Adults $\geq$ Bachelor Deg.	0.059^{**}
$\log_2$(Median Home Value)	0.010^{*}
$\log_2$(Number of students)	0.006^{*}
$\log_2$(Student:Teacher ratio)	-0.007
Is Public	0.015^{*}
Is Magnet	0.013
Is Charter	0.033
R^2	0.140
adj. R^2	0.133

Table 2: Regression of the average P(high quality) of a school on demographic variables ($N = 968$). We observe that larger schools in educated, urban, and wealthy areas of the U.S tend to be scored higher by the GPT-3 quality filter. See §A.8 for more information on these features. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

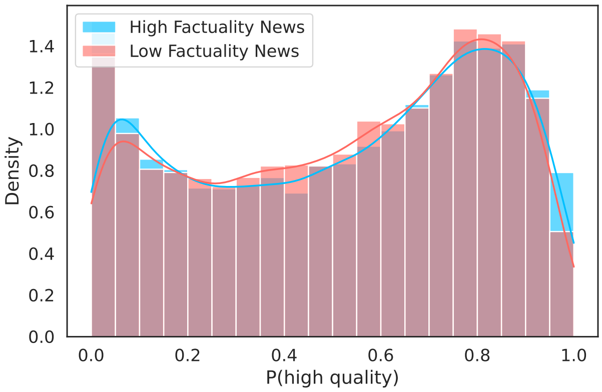

Figure 2: There is no difference in quality scores between articles written by news sources of high and low factual reliability.

Nevertheless, the effects, considered individually, are relatively modest. A 14 percentage point increase in percent urban population or a 17 percentage point increase in parental education (percent of adults with college degrees) correspond to a 1 percentage point increase in average quality score, as does a doubling of home values, or a quadrupling of school size (holding other variables constant in each case). Average quality scores associated with public schools are 1.5 percentage points higher than private schools, controlling for other factors. Coefficients for charter schools, magnet schools, and student:teacher ratio are not significant. The combined effects of all these factors account for large differences in quality scores between wealthy, urban, educated locations, and poorer, rural, and less educated parts of the United States.

Summary and Limitations This analysis reveals an unintended consequence of the GPT-3 quality filter: by attempting to exclude text that is less like mainstream news and Wikipedia, the filter reinforces a language ideology that text from authors of wealthy, urban, and educated backgrounds is more valuable for inclusion in language model training data. These implicit preferences align with the attributes of authors that dominate the corpora from §2, which the filter considers to be high quality.

While most of the above findings are robust to alternate model specifications, the model ultimately only accounts for a relatively small amount of variance in quality scores. However, given that all variation is ultimately explained by features of text itself, any amount of variance accounted for by demographic features is notable.

In addition, most of our features are taken from a single point in time and do not account for changing demographics over the examined time period (2010–2019). Data errors could also arise due to how datasets were aligned (based on school name and ZIP code). These findings may not generalize to other domains (e.g., social media), and inclusion of additional features could affect these findings. For additional models which include party vote share and racial demographics taken from NCES data, see §A.9.

4 Alignment with Other Notions of Quality

The GPT-3 quality filter purports to judge the quality of text, something that people also do, using a variety of different criteria. In this section, we consider three forms of human evaluations: factuality judgements, human-graded standardized test essays, and institutional book awards. How well does the behavior of the GPT-3 quality filter map onto these notions of quality?

4.1 Data

Factually (Un)reliable News To analyze the correspondence between the GPT-3 quality filter and news factuality, we use the list provided by Baly et al. (2018) to identify a set of popular news sources from a broad range of factuality ratings and political leanings.[17] Using `Newspaper3k`,[18]

[17] Baly et al. (2018) release a dataset of factual reliability and political leanings across news sources by scraping `NewsMediaBiasFactCheck.org`.

[18] `https://newspaper.readthedocs.io/`

we scrape and score 9.9K and 7.7K articles from high and low factuality news outlets, respectively.

TOEFL Essay Exams Next, to analyze the correspondence between the GPT-3 quality filter and essay scores, we collect and score 12.1K participant essays from the *Test Of English as a Foreign Language* (TOEFL) exam, a widely used English language proficiency test (Blanchard et al., 2013). The TOEFL exam responses include official scores from exam readers, as well as each essay's prompt.

Award-Winning Literature Finally, to analyze the correspondence between the GPT-3 quality filter and literary awards, we select and score books from Books3 and the Gutenberg corpus (Brooke et al., 2015) that have won a Pulitzer Prize in various categories. We collected these data by scraping the publicly available list of award recipients.[19]

4.2 Results

If the filter aligns with news factuality, we would expect that articles from factually reliable sources would be rated as higher quality than those from factually unreliable ones. However, we find no difference in the quality distribution between articles from high and low factuality news sources ($p = 0.085$, two-way Kolmogorov-Smirnov test; Figure 2). Many factually unreliable news articles are considered high quality by the filter (§A.10).

Turning to the TOEFL exam responses, we would expect that if the filter agrees with essay scores, higher scoring essays would receive higher quality scores. While essay scores are weakly correlated with quality scores (Pearson $r = 0.12$, $p < 0.001$), the essay's prompt is far more predictive of the essay's quality designation (§A.11). For example, essays responding to a prompt (*P4*) which asks participants to describe *"...whether advertisements make products seem much better than they really are"* are much less likely to be filtered than all other prompts, including *P6*, which asks participants to describe *"...whether it is best to travel in a group"* (see §A.11 for more details). The latter prompt tends to invoke personal experiences in the responses.

Finally, if the filter aligns with literary awards, we would expect that most Pulitzer-Prize winning books would achieve high quality scores. On the contrary, quality scores vary heavily based on the genre (Figure 3). Poetry and drama are less favored by the filter, relative to non-fiction, fiction, and fan fiction (from BookCorpus; Zhu et al. 2015).

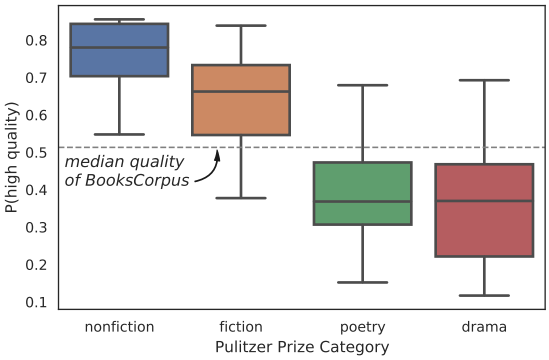

Figure 3: Among works that have won a Pulitzer Prize, the quality filter tends to favor nonfiction and longer fictional forms, disfavoring poetry and dramatic plays.

Summary Our analysis demonstrates that the GPT-3 quality filter conflicts with other standards of text quality. Of course, even the alternative standards we compare here are subject to their own language ideologies. Readers are more likely to trust news as factual if its political position aligns with their own (Mitchell et al., 2018). English-language teaching pedagogies are rooted in ideologies about well-spokenness (Vanegas et al., 2016). Literary awards favor white and male authors.[20] In general, any designation of text as high quality is subjective and influenced by sociopolitical context.

5 Discussion

The above sections have demonstrated that automated filtering of text to build language modeling corpora may lead to counterintuitive or undesirable exclusion of sources. Because of the variety of use cases for language models and the broad range of text that could be appropriate for certain tasks, we suggest that there is no simple, universal standard for what should be considered high quality text. Indeed, there is a long history of privileging some people's spoken language as better or more "correct" than others. Researchers and practitioners of NLP who are aware of this history have the option to be intentional in their design of systems that, however implicitly, risk excluding the language of underprivileged identities or communities.

[19] https://www.pulitzer.org/prize-winners-categories

[20] A 2016 study by the Columbia Journalism Review found that since 1918, 84% of Pulitzer Prizes had been awarded to white authors, and 84% to male authors: https://www.cjr.org/analysis/100_years_of_data.php.

Some amount of selection in building corpora is inevitable. It is not possible to collect a uniform random sample of all written utterances. However, our findings suggest that current selection methods are, for many purposes, flawed. Future work into alternative filtering criteria could be paired with investigations into the unintended consequences of their assumptions.

We do not believe that there is likely to be a single solution to this challenge. Indeed, the text that is best suited for training a model may depend on the application of that model. At a minimum, however, the NLP community could more carefully consider and clearly document the inclusion criteria for text. NLP practitioners could also be explicit about their reasons for using certain sources, even if those reasons are related to availability or empirical performance. A collection of tests could also be deployed (and improved over time) to give a clear understanding of the implications of different choices of filters.

More generally, we echo calls in the literature for more thoughtful and inclusive data collection (Jo and Gebru, 2020; Bender et al., 2021; Tanweer et al., 2021). Strategies could include, but are not limited to a) intentionally curating data from people and viewpoints that are not otherwise well represented; b) including a greater diversity of genres; c) adopting more nuanced or intentional exclusion criteria; d) conducting more thorough interrogation of what text is being excluded; e) developing standard checks for prominent biases in inclusion; and/or f) abandoning the notion of a general-purpose corpus.

6 Related Work

Language Ideologies Language ideologies have been widely explored in the sociolinguistics literature (Gal and Irvine, 1995; Rosa and Flores, 2017; Craft et al., 2020, *inter alia*). An ideology that promotes the inherent correctness, clarity, and objectivity of certain language varieties over others is a mechanism for linguistic discrimination (Craft et al., 2020; Gal, 2016; MacSwan, 2020; Rickford and King, 2016). A salient example of such discrimination is the stigmatization of second-language speakers of English (Lindemann, 2005).

Language ideologies have an important, but often unacknowledged, influence on the development of NLP technologies (Blodgett et al., 2020). For example, an ideology that distinguishes between *standard* and *non-standard* language variations surfaces in text normalization tasks (van der Goot et al., 2021), which tend to strip documents of pragmatic nuance (Baldwin and Chai, 2011) and social signals (Nguyen et al., 2021). Language on the Internet has been historically treated as a noisy variant of English, even though lexical variation on the Internet is highly communicative of social signals (Eisenstein, 2013) and varies considerably along demographic variables (Eisenstein et al., 2014) and community membership (Lucy and Bamman, 2021).

Language ideologies also surface in tools for toxicity detection; for example, the classification behavior of the PERSPECTIVE API (a popular hate speech detector) aligns with the attitudes of conservative, white, female annotators, who tend to perceive African-American dialects as more toxic (Sap et al., 2021).

Critiques of *Laissez-Faire* Data Collection We provide empirical evidence that *laissez-faire* data collection (i.e., filtering large web data sources) leads to data homogeneity (Bender et al., 2021). As an alternative to *laissez-faire* collection, Jo and Gebru (2020) recommend drawing on institutional archival practices. However, we note that language ideologies are also prevalent (and may not be explicit) in institutional archives, which, for example, have preferred colonizing perspectives over colonized ones when documenting historical events (Trouillot, 1995; Decker, 2013).

Other Quality Filters Other definitions of text quality are used to create pretraining datasets, some of which do not rely on the datasets from §2. However, all techniques adopt language ideologies of what constitutes high quality text. *Bad-word* filtering, which removes documents that contain certain stop-words, disproportionately excludes language about and by non-dominant groups (Dodge et al., 2021). Filtering Internet content for popularity (Radford et al., 2019) leads to data homogeneity based on the characteristics of viral media and the composition of userbases in online forums (§2). Even lightweight filters (Aghajanyan et al., 2021; Rae et al., 2021) put more emphasis on features like document length over factuality when determining what makes a document high quality. Any filtering method requires transparent justification and recognition of tradeoffs.

Downstream Behavior The behavior of language processing systems aligns with what we would expect from a language ideology that favors training data written by a narrow, powerful sector of society. For example, dialogue agents perform significantly worse when engaging in conversations about race (Schlesinger et al., 2018) and with non-dominant dialects of English (Mengesha et al., 2021). GPT-3 frequently resorts to using stereotypes when minority groups are mentioned in its prompt (Abid et al., 2021; Blodgett, 2021). GPT-3 is also prone to producing hate speech (Gehman et al., 2020) and misinformation (McGuffie and Newhouse, 2020), which we would expect if its quality filter fails to distinguish the factual reliability of news sources in its training data (§4). Gao (2021) show that aggressive data filtering with the GPT-3 quality filter degrades downstream task performance. A closer analysis of how the language ideologies in data selection lead to certain model behaviors is a rich area for future work.

7 Conclusion

Using a new dataset of U.S. school newspapers, we find that the conventional, automated valuation of Wikipedia, newswire, books, and popular Internet content as reference for high quality text implicitly favors content written by authors from larger schools in wealthier, educated, urban areas of the United States. Adopting this language ideology for text data selection leads to implicit—yet systematic and as-yet undocumented—inequalities in terms of whose language is more likely to be included in training corpora. Although no single action will solve this complicated issue, data curators and researchers could be more intentional about curating text from underrepresented authors and groups, gathering sources from multiple genres and writing styles, and documenting their curation procedures and possible sources of exclusion.

Ethical Considerations

Our U.S. SCHOOL NEWS dataset comes with many limitations, as described in §3.1. Our corpus is neither a random nor a representative sample of U.S. school newspapers. Instead, it represents schools that had sufficient Internet access, that elected to use a particular website template, and that maintain websites with retrievable archived content. In general, our dataset likely captures neither the least resourced schools (which may not have good access to online resources) in the United States, nor the wealthiest ones (who may have their own publication platforms). The lack of representation in school newspaper leadership positions may influence which students contribute content to school newspapers (Chen et al., 2021). Educators also likely shape some articles, at least in part (though we expect them to be similarly affected by resource constraints).

Moreover, much of the content in these articles is specific to student concerns (e.g., sports, school events, campus culture, etc.), and the writing is, by definition, amateur. Nevertheless, because the corpus captures a wide range of content and geographical areas, it allows us to evaluate how a quality filter handles real-world language variation within a particular domain. Additionally, we speculate that an expanded corpus, which included writings from these schools, would demonstrate a continuation of trends we report in this paper.

Using text from school newspapers introduces privacy concerns, especially since authors and subjects are minors. We therefore use this data only for evaluation purposes; we do **not** train (or release) any models on this data or on any raw text from the corpus. We do, however, release a datasheet (Gebru et al., 2021) which documents the dataset's general characteristics and curation procedure (§A.3).

While the text in our dataset varies considerably along topical, stylistic, and demographic variables, it is nevertheless a niche domain. The text is a specific genre meant for local student consumption, its authors are U.S. students, and it thus primarily represents U.S.-centric cultural and political perspectives. We acknowledge that we also perpetuate some of the biases we identify, especially by working with English language text from the United States. We hope future work will extend this study of language ideologies to multilingual settings, other textual domains, and different sets of authors.

With respect to demographic variables, we merge census demographics with school-level data via ZIP codes or counties, which are imperfect identifiers of a school, since ZIP codes (and counties) may include multiple schools of varying resource levels. Moreover, tracking demographic variables and other author metadata, if deployed at scale, implies a certain level of invasive surveillance (Brayne, 2017). Future work may explore how to maintain the rights of authors as data subjects and producers while mapping demographic representation in large corpora.

The lack of access to GPT-3's training data and quality filter prevents us from making claims about how quality filter biases affect language model behavior. Future work on language models may also include transparent release of training data and associated quality filters, which would help support this kind of research.

Finally, we did not seek consent from authors to scrape their articles. The ethical and legal norms around scraping public-facing web data, especially those produced by minors, are still in flux (Fiesler et al., 2020) and may not align with user perceptions of what constitutes fair use of online communications (Williams et al., 2017). For these reasons (as discussed earlier), we do not release the corpus of school newspaper articles, and only use it for analysis and evaluation. We only make available a dataset of demographic variables and quality scores per school, to support reproducibility.

Acknowledgments

This paper benefited from thoughtful feedback from a number of people: Emily M. Bender, Su Lin Blodgett, Amber Boydstun, Timnit Gebru, Eun Seo Jo, Kelvin Luu, Lucy Li, Julian Michael, Amandalynne Paullada, Katharina Reinecke, Swabha Swayamdipta, Kelly E. Wright, and Kaitlyn Zhou.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-Muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*.

Armen Aghajanyan, Dmytro Okhonko, Mike Lewis, Mandar Joshi, Hu Xu, Gargi Ghosh, and Luke Zettlemoyer. 2021. Htlm: Hyper-text pre-training and prompting of language models. *arXiv*, abs/2107.06955.

Gabriel Arana. 2018. Decades of failure. *Columbia Journalism Review*.

Tyler Baldwin and Joyce Chai. 2011. Beyond normalization: Pragmatics of word form in text messages. In *Proceedings of 5th International Joint Conference on Natural Language Processing*.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of EMNLP*.

David Bamman and Noah A Smith. 2014. Unsupervised discovery of biographical structure from text. *Transactions of the Association for Computational Linguistics*, 2:363–376.

Jack Bandy and Nicholas Vincent. 2021. Addressing "documentation debt" in machine learning: A retrospective datasheet for BookCorpus. In *NeurIPS*.

Michael Barthel, Galen Stocking, Jesse Holcomb, and Amy Mitchell. 2016. Seven-in-Ten Reddit users get news on the site. [online; accessed: 2020-6-2].

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of FAccT*.

Julian R. Betts, Kim S. Reuben, and Anne Danenberg. 2000. *Equal Resources, Equal Outcomes? The Distribution of School Resources and Student Achievement in California*. Public Policy Institute of California.

Steve Bien-Aimé. 2016. AP stylebook normalizes sports as a male space. *Newspaper Research Journal*, 37(1):44–57.

Daniel Blanchard, Joel R. Tetreault, Derrick Higgins, A. Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native English. *ETS Research Report Series*, 2013:15.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Su Lin Blodgett. 2021. *Sociolinguistically Driven Approaches for Just Natural Language Processing*. Ph.D. thesis, University of Massachusetts Amherst.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of EMNLP*.

Sarah Brayne. 2017. Big data surveillance: The case of policing. *American Sociological Review*, 82(5):977–1008.

Julian Brooke, Adam Hammond, and Graeme Hirst. 2015. GutenTag: an NLP-driven tool for digital humanities research in the Project Gutenberg corpus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric

Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *arXiv*, abs/2005.14165.

Janice Kai Chen, Ilena Peng, Jasen Lo, Trisha Ahmed, Simon J. Levien, and Devan Karp. 2021. Voices investigation: Few black, latinx students are editors of top college newspapers. *AAJA Voices*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Justin T. Craft, Kelly E. Wright, Rachel Elizabeth Weissler, and Robin M. Queen. 2020. Language and discrimination: Generating meaning, perceiving identities, and discriminating outcomes. *Annual Review of Linguistics*, 6(1):389–407.

Stephanie Decker. 2013. The silence of the archives: business history, post-colonialism and archival ethnography. *Management & Organizational History*, 8(2):155–173.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Robert DiNicola. 1994. Teaching journalistic style with the AP stylebook: Beyond fussy rules and dogma of 'correctness'. *The Journalism Educator*, 49(2):64–70.

Jesse Dodge, Maarten Sap, Ana Marasovic, William Agnew, Gabriel Ilharco, Dirk Groeneveld, and Matt Gardner. 2021. Documenting the English Colossal Clean Crawled Corpus. *arXiv*, abs/2104.08758.

Penelope Eckert. 1989. *Jocks and burnouts: Social categories and identity in the high school*. Teachers college press.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of NAACL*, pages 359–369.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLoS ONE*, 9.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity.

Casey Fiesler, Nathan Beard, and Brian Keegan. 2020. No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service. In *Proceedings of ICWSM*.

Paula Froke, Anna Jo Bratton, Jeff McMillan, Pia Sarkar, Jerry Schwartz, and Raghuram Vadarevu. 2020. *The Associated Press stylebook 2020-2022*. The Associated Press.

Susan Gal. 2016. *Sociolinguistic differentiation*, page 113–136. Cambridge University Press.

Susan Gal and Judith T. Irvine. 1995. The boundaries of languages and disciplines: How ideologies construct difference. *Social Research*, 62(4):967–1001.

Leo Gao. 2021. An empirical exploration in quality filtering of text data. *arXiv*, abs/2109.00698.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800Gb dataset of diverse text for language modeling. *arXiv*, abs/2101.00027.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Joyce Still Gibson. 1961. A study of the status of high school newspapers in the virginia public schools. Master's thesis, University of Richmond.

Aaron Gokaslan and Vanya Cohen. 2019. OpenWebText corpus.

Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. 2015. First women, second sex: Gender bias in Wikipedia. In *Proceedings of the 26th ACM conference on hypertext & social media*.

Rob Greenwald, Larry V. Hedges, and Richard D. Laine. 1996. The effect of school resources on student achievement. *Review of Educational Research*, 66(3):361–396.

Elizabeth Grieco. 2018. Newsroom employees are less diverse than U.S. workers overall. *Pew Research Center*. [online; accessed 2022-01-22].

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of NAACL*.

Keira Huang. 2013. Wikipedia fails to bridge gender gap. *South China Morning Post*. [online; accessed 2022-01-11].

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. *Proceedings of FAccT*.

Paresh Kharya and Ali Alvi. 2021. Using DeepSpeed and Megatron to train Megatron-Turing NLG 530b, the world's largest and most powerful generative language model. [online; accessed 2022-01-20].

William Labov. 2006. *The Social Stratification of English in New York City*, 2 edition. Cambridge University Press.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Lee & Low Books. 2020. Where is the diversity in publishing? The 2019 diversity baseline survey results. [online; accessed 2021-11-24].

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Stephanie Lindemann. 2005. Who speaks "broken English"? US undergraduates' perceptions of non-native English. *International Journal of Applied Linguistics*, 15(2):187–212.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv*, abs/1907.11692.

Li Lucy and David Bamman. 2021. Characterizing English Variation across Social Media Communities with BERT. *Transactions of the Association for Computational Linguistics*, 9:538–556.

Jeff MacSwan. 2020. Academic English as standard language ideology: A renewed research agenda for asset-based language education. *Language Teaching Research*, 24(1):28–36.

Michael Mandiberg. 2020. Mapping Wikipedia. *The Atlantic*. [online; accessed 2021-11-24].

Sorin Adam Matei and Brian C. Britt. 2017. *Structural Differentiation in Social Media*. Springer International Publishing.

Kris McGuffie and Alex Newhouse. 2020. The radicalization risks of GPT-3 and advanced neural language models. *arXiv*, abs/2009.06807.

Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. 2021. "I don't think these devices are very culturally sensitive."—Impact of automated speech recognition errors on African Americans. *Frontiers in Artificial Intelligence*, 4:169.

Meta-wiki. 2018. Community insights/2018 report/contributors. [online; accessed 2012-11-24].

Amy Mitchell, Jeffrey Gottfried, Michael Barthel, and Nami Sumida. 2018. Can Americans tell factual from opinion statements in the news? *Pew Research Center's Journalism Project*. [online; accessed 2022-01-22].

Dong Nguyen, Laura Rosseel, and Jack Grieve. 2021. On learning and representing social meaning in NLP: A sociolinguistic perspective. In *Proceedings of NAACL*.

Katherine Panciera, Aaron Halfaker, and Loren Terveen. 2009. Wikipedians are born, not made: A study of power editors on Wikipedia. In *Proceedings of the ACM 2009 International Conference on Supporting Group Work*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. [online; accessed 2022-01-22].

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. [online; accessed 2022-01-22].

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv*, abs/2112.11446.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sean F Reardon and Ann Owens. 2014. 60 years after Brown: Trends and consequences of school segregation. *Annual Review of Sociology*, 40:199–218.

John R. Rickford. 1985. Ethnicity as a sociolinguistic boundary. *American Speech*, 60(2):99–125.

John R. Rickford and Sharese King. 2016. Language and linguistics on trial: Hearing Rachel Jeantel (and other vernacular speakers) in the courtroom and beyond. *Language*, 92(4):948–988.

Jonathan Rosa and Nelson Flores. 2017. Unsettling race and language: Toward a raciolinguistic perspective. *Language in Society*, 46(5):621–647.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv*, abs/2111.07997.

Dante J. Scala and Kenneth M. Johnson. 2017. Political polarization along the rural-urban continuum? the geography of the presidential vote, 2000–2016. *The ANNALS of the American Academy of Political and Social Science*, 672(1):162–184.

Ari Schlesinger, Kenton P. O'Hara, and Alex S. Taylor. 2018. Let's talk about race: Identity, chatbots, and AI. In *Proceedings of CHI*.

Anissa Tanweer, Emily Kalah Gade, PM Krafft, and Sarah K Dreier. 2021. Why the data revolution needs qualitative thinking. *Harvard Data Science Review*.

Michel-Rolph Trouillot. 1995. *Silencing the past: Power and the production of history*. Beacon Press.

Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešić, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoğlu, Timothy Baldwin, Tommaso Caselli, and Wladimir Sidorenko. 2021. MultiLexNorm: A shared task on multilingual lexical normalization. In *Proceedings of the Seventh Workshop on Noisy User-generated Text*.

Marlon Vanegas, Juan Restrepo, Yurley Zapata, Giovany Rodríguez, Luis Cardona, and Cristian Muñoz. 2016. Linguistic discrimination in an English language teaching program: Voices of the invisible others. *Íkala, Revista de Lenguaje y Cultura*, 21.

Fred Vultee. 2012. A paleontology of style. *Journalism Practice*, 6(4):450–464.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Proceedings of the AAAI conference on web and social media*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of LREC*.

Matthew L Williams, Pete Burnap, and Luke Sloan. 2017. Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, 51(6):1149–1168.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of ICCV*.

A Appendix

A.1 Language Model Training Corpora

We display a list of popular language modeling corpora in Table 3.

A.2 OpenWebText URL distribution

We display the most popular URL domains of OpenWebText in Table 4.

A.3 Datasheet

Our datasheet for the U.S. SCHOOL NEWS dataset can be found here: `https://bit.ly/3tSpYt8`.

A.4 Quality Filter Hyperparameters

We display the hyperparameters of our logistic regression classifier (reproduction of the filter developed by Brown et al. 2020) in Table 5.

A.5 High School News Scores

We display the score distributions of school articles in our U.S. SCHOOL NEWS dataset, relative to general newswire, in Figure 4.

A.6 Example Articles

We display example articles and their quality scores in the U.S. SCHOOL NEWS dataset in Table 6.

A.7 Topic Modeling

See the quality distribution among topics for 10K opinion pieces in Figure 5.

A.8 Demographic Features

We display a table of features we use in our demographic regression model in Table 7.

A.9 Additional Regressions

Here we include regressions results from two models with additional covariates.

We first consider race as a possible omitted variable, given the extent of school segregation in the U.S. (Reardon and Owens, 2014). NCES data provides the distribution of students by race for each school, using a particular set of racial categories, which comes with obvious limitations. Nevertheless, we use the raw percentage scores provided as additional covariates in this model as a validity check. We exclude the Native and Pacific Islander categories, due to imbalanced data and geographic concentration, as well as the white category, to avoid a saturated model.

As shown in Table 8, the findings are nearly identical to the results in the main paper, with the exception that home values are no longer significant. The only racial category that shows a significant effect is Asian. However, we note a positive correlation between percentage of Asian students and median home values (Pearson r =0.32, $p < 0.001$), suggesting that the variable for percentage of Asian students may be partially absorbing the effect of our measure of wealth.

Table 9 shows the results for an alternate model which includes % GOP vote share in the 2016 election. Once again, the results are very similar to the results in the main paper, although there is a strong (and significant) negative association between GOP vote share and quality scores, whereas the measures of home values and percent rural are no longer significant.

The results for this model exemplify the difficulty of working with highly correlated variables. Given the strong association between GOP voters and rural areas, GOP vote share serves as an effective proxy for other variables of interest. However, because the results of the 2016 Presidential election were likely somewhat idiosyncratic, and because we find wealth and geography to be a more plausible explanation for differences in student writing than political preferences among their parents, we opt for the model without GOP vote share in the main paper.

A.10 Low Factuality News Considered High Quality

We display example low factuality news articles that are assigned high quality scores by the GPT-3 quality filter in Table 10.

A.11 TOEFL Exam Responses

We display a regression of the quality of a TOEFL exam essay on its assigned score and prompt in Table 11. We display the distribution of quality scores against prompts and essay scores in the TOEFL exam dataset in Figure 6. We display the prompts of this dataset in Table 12.

Model	Pretraining Data Sources	Citation
ELMo	1B Word benchmark	Peters et al. 2018
GPT-1	BookCorpus	Radford et al. 2018
GPT-2	WebText	Radford et al. 2019
BERT	BookCorpus + Wikipedia	Devlin et al. 2019
RoBERTa	BookCorpus + Wikipedia + CC-news + OpenWebText + Stories	Liu et al. 2019
XL-Net	BookCorpus + Wikipedia + Giga5 + ClueWeb 2012-B + Common Crawl	Yang et al. 2019
ALBERT	BERT, RoBERTa, and XL-net's data sources	Lan et al. 2020
T5	Common Crawl (filtered)	Raffel et al. 2020
XLM-R	Common Crawl (filtered)	Conneau et al. 2020
BART	BookCorpus + Wikipedia	Lewis et al. 2020
GPT-3	Wikipedia + Books + WebText (expanded) + Common Crawl (filtered)	Brown et al. 2020
ELECTRA	BookCorpus + Wikipedia + Giga5 + ClueWeb 2012-B + Common Crawl	Clark et al. 2020
Megatron-Turing NLG	The Pile + Common Crawl (filtered) + RealNews + Stories	Kharya and Alvi 2021
Switch-C	Common Crawl (filtered)	Fedus et al. 2021
Gopher	MassiveWeb + Books + Common Crawl (filtered) + News + GitHub + Wikipedia	Rae et al. 2021

Table 3: Overview of recent language models and their training corpora. All studies tend to draw from the same core data sources: Wikipedia, Books, News, or filtered web dumps.

URL Domain	# Docs	% of Total Docs
bbc.co.uk	116K	1.50%
theguardian.com	115K	1.50%
washingtonpost.com	89K	1.20%
nytimes.com	88K	1.10%
reuters.com	79K	1.10%
huffingtonpost.com	72K	0.96%
cnn.com	70K	0.93%
cbc.ca	67K	0.89%
dailymail.co.uk	58K	0.77%
go.com	48K	0.63%

Table 4: The most popular top-level URL domains in OpenWebText. Mainstream news forms the overwhelming majority of content in the dataset. Overall, just 1% of the top-level URL domains in OpenWebText contribute 75% of the total documents in the corpus.

Computing Infrastructure	56 Intel Xeon CPU Cores
Number of search trials	100
Search strategy	uniform sampling
Best validation F1	90.4

Hyperparameter	Search space	Best assignment
regularization	*choice*[L1, L2]	L1
C	*uniform-float*[0, 1]	0.977778
solver	64	liblinear
tol	*loguniform-float*[10e-5, 10e-3]	0.000816
ngram range	*choice*["1 2", "1 3", "2 3"]	"1 2"
random state	*uniform-int*[0, 100000]	44555
tokenization	whitespace	whitespace
vectorization	hashing	hashing
remove stopwords	*choice*[Yes, No]	No

Table 5: Hyperparameter search space and best assignments for our re-implementation of the GPT-3 quality filter.

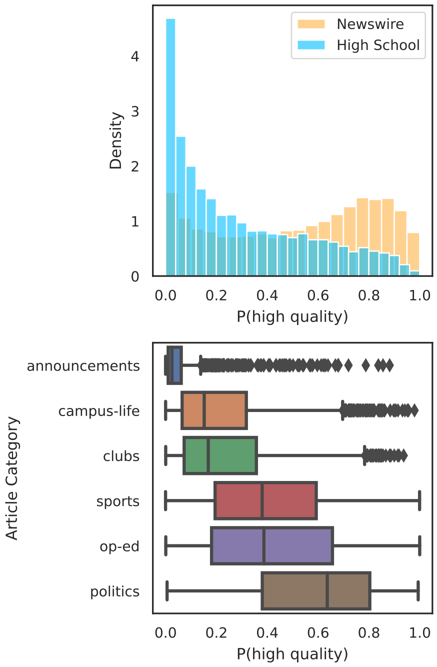

Figure 4: Scraped school articles tend to be considered lower quality by the GPT-3 quality filter than general newswire (histogram built from 10K random documents from each domain). This finding is consistent across a variety of categories, and more significant for certain ones (e.g., school announcements).

Category: Student-Life P**(high quality) = 0.001**
As our seniors count down their final days until graduation, we will be featuring them each day. [REDACTED], what are your plans after graduation? To attend [REDACTED] in the fall and get my basics. Then attend the [REDACTED] program. What is your favorite high school memory? My crazy, obnoxious and silly 5th hour English with [REDACTED]. What advice do you have for underclassmen? Pay attention, stay awake (I suggest lots of coffee), and turn in your dang work! You can do it, keep your head up because you are almost there!
Category: News P**(high quality) = 0.99**
On Monday, September 3rd, Colin Kaepernick, the American football star who started the "take a knee" national anthem protest against police brutality and racial inequality, was named the new face of Nike's "Just Do It" 30th-anniversary campaign. Shortly after, social media exploded with both positive and negative feedback from people all over the United States. As football season ramps back up, this advertisement and the message behind it keeps the NFL Anthem kneeling protest in the spotlight.

Table 6: Examples of high school news paper articles from U.S. SCHOOL NEWS. Many of the articles in student-life category, and similar, rated lower quality have very different styles from documents rated high quality.

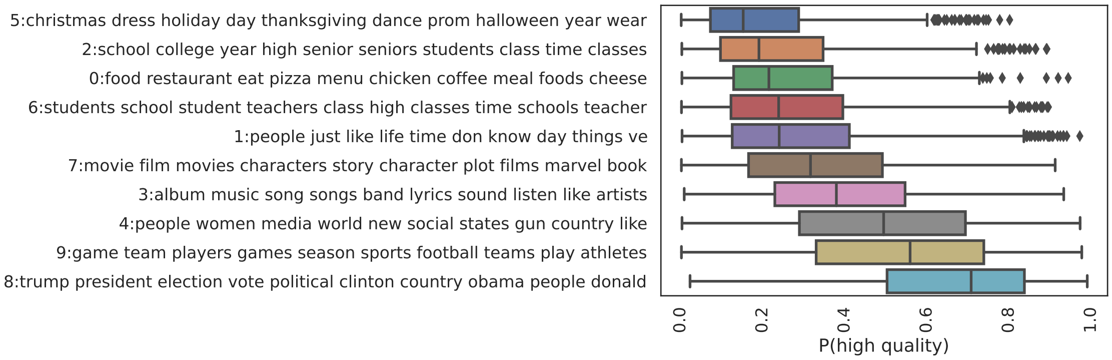

Figure 5: Considering 10K opinion pieces in U.S. SCHOOL NEWS, we observe that the GPT-3 quality filter prefers topics that are more prevalent in Wikipedia or newswire.

Feature	Description	Level	Source
Is Charter	Is the school a charter school?	School	NCES database
Is Private	Is the school a private school?	School	NCES database
Is Magnet	Is the school a magnet school?	School	NCES database
% Black Students	% students who identify as Black	School	NCES database
% Asian Students	% students who identify as Asian	School	NCES database
% Mixed Students	% students who identify as Mixed race	School	NCES database
% Hispanic Students	% students who identify as Hispanic	School	NCES database
Student:Teacher	Student-teacher ratio	School	NCES database
School Size	Total number of students	School	NCES database
Median Home Value	Median home value	ZIP code	Census
% Adults $\geq$ Bachelor Deg.	% adults ($\geq$ 25 years old) with at least a bachelor's degree	ZIP code	Census
% Rural	Percent of a county population living in a rural area	County	Census
% 2016 GOP Vote	Republican vote share in the 2016 presidential election	County	MIT Election Lab

Table 7: Description of features we include in our demographic analyses.

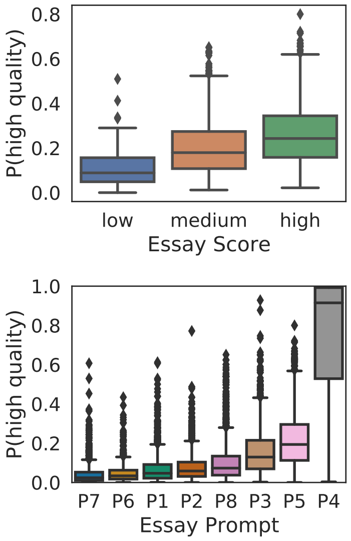

Figure 6: TOEFL exam score is weakly correlated with quality score across prompts (Pearson correlation; r=0.12 $\pm$ 0.05, $p \approx 0$; top), but the essay prompt seems to be a much stronger indicator of quality scores than the exam scores are (bottom).

Dependent variable: P(high quality)
Observations: 968 schools

Feature	Coefficient
Intercept	0.134
% Rural	-0.073^{***}
% Adults $\geq$ Bachelor Deg.	0.049^{*}
$\log_2$(Median Home Value)	0.007
$\log_2$(Number of students)	0.005^{*}
$\log_2$(Student:Teacher ratio)	-0.008
Is Public	0.020^{*}
Is Magnet	0.013
Is Charter	0.035^{*}
% Asian Students	0.081^{**}
% Mixed Students	0.051
% Black Students	-0.009
% Hispanic Students	-0.020
R^2	0.152
adj. R^2	0.142

Table 8: Regression of the average P(high quality) of a school in the U.S. SCHOOL NEWS dataset, on demographic variables. As in the main paper, larger schools in educated and urban areas of the U.S tend to be scored higher by the GPT-3 quality filter. Asian is the only categorical race variable which shows a significant association (using data and categories taken directly from NCES). The association with home values is no longer significant, plausibly explained by a correlation between a higher proportion of Asian students and higher median home values. See §A.8 for more information on these features. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Dependent variable: P(high quality)
Observations: 968 schools

Feature	Coefficient
Intercept	0.248^{**}
% Rural	-0.021
% Adults $\geq$ Bachelor Deg.	0.067^{**}
$\log_2$(Median Home Value)	0.003
$\log_2$(Number of students)	0.006^{**}
$\log_2$(Student:Teacher ratio)	-0.007
Is Public	0.017^{*}
Is Magnet	0.009
Is Charter	0.027
% GOP vote share	-0.114^{***}
R^2	0.164
adj. R^2	0.157

Table 9: Regression of the average P(high quality) of a school in the U.S. SCHOOL NEWS dataset, on demographic variables, including % 2016 GOP Vote. We observe that including the political leaning of the county tends to wash out other variables, likely because partisan voting correlates heavily with other effects, like the urban/rural divide (Scala and Johnson, 2017). The only other covariates that stay significant are school size, parental education, and public (as opposed to private) schools. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Article from *http://en-volve.com* **P(high quality) = 0.93**
The German government has effectively began the process of eliminating the unvaccinated by starving them to death by pushing grocery stories to ban unvaccinated residents from buying essential food items...The pressure on the unvaccinated grows and grows!...
Article from *http://www.censored.news* **P(high quality) = 0.98**
The provisional number of births in the U.S. was 3,605,201 in 2020. That is the lowest number of births in the United States since 1979, according to the Centers for Disease Control. 2020 also had the lowest fertility rate since the government started tracking births in 1902. And don't blame the so-called "pandemic."...we're learning in 2021 that intelligent people succumb to government psy-ops. But critical thinkers understood immediately that something was very wrong with all the COVID-19 stuff. Plus many among the global elite continually and openly gloat about their desire to cull the masses. Bill Gates isn't even coy about his desires...

Table 10: Examples of news from low factuality sources (as identified by `MediaBiasFactCheck.com`) rated high quality by GPT-3 quality filter, but contain COVID disinformation.

Dependent variable: P(high quality)
Observations: 12.1K TOEFL exams

Feature	Coefficient
Intercept	0.0631^{***}
Low score	-0.0414
High score	0.0339
Prompt 7	-0.0283^{***}
Prompt 6	-0.0204^{***}
Prompt 2	0.0068^{***}
Prompt 8	0.0346^{***}
Prompt 3	0.0880^{***}
Prompt 5	0.1470^{***}
Prompt 4	0.6745^{***}
R^2	0.712
adj. R^2	0.711

Table 11: Regression of the quality of a TOEFL exam essay on its assigned score and prompt. While we observe some relationship between the score an essay receives and its quality score, the essay prompts themselves have significantly higher effect sizes. The highest quality essays come from Prompt 4, which asks participants to discuss products and advertisements. See §A.11 for visualizations of distributions of quality across prompts and scores. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

ID	Text	P(**high quality**)
P7	It is more important for students to understand ideas and concepts than it is for them to learn facts.	0.04
P6	The best way to travel is in a group led by a tour guide.	0.05
P1	It is better to have broad knowledge of many academic subjects than to specialize in one specific subject.	0.07
P2	Young people enjoy life more than older people do.	0.08
P8	Successful people try new things and take risks rather than only doing what they already know how to do well.	0.10
P3	Young people nowadays do not give enough time to helping their communities.	0.16
P5	In twenty years, there will be fewer cars in use than there are today.	0.22
P4	Most advertisements make products seem much better than they really are.	0.74

Table 12: TOEFL prompt IDs and their text, ordered by their quality score by GPT-3 quality filter.