AboutMe: Using Self-Descriptions in Webpages to Document the Effects of English Pretraining Data Filters

Li Lucy[1,2] **Suchin Gururangan**[5] **Luca Soldaini**[1]
Emma Strubell[1,4] **David Bamman**[2] **Lauren F. Klein**[3] **Jesse Dodge**[1]
[1]Allen Institute for AI [2]University of California, Berkeley [3]Emory University
[4]Carnegie Mellon University [5]University of Washington
lucy3_li@berkeley.edu

Abstract

Large language models' (LLMs) abilities are drawn from their pretraining data, and model development begins with data curation. However, decisions around what data is retained or removed during this initial stage are under-scrutinized. In our work, we ground web text, which is a popular pretraining data source, to its social and geographic contexts. We create a new dataset of 10.3 million self-descriptions of website creators, and extract information about who they are and where they are from: their topical interests, social roles, and geographic affiliations. Then, we conduct the first study investigating how ten "quality" and English language identification (langID) filters affect webpages that vary along these social dimensions. Our experiments illuminate a range of implicit preferences in data curation: we show that some quality classifiers act like topical domain filters, and langID can overlook English content from some regions of the world. Overall, we hope that our work will encourage a new line of research on pretraining data curation practices and its social implications.

1 Introduction

Large language models (LLMs) are sometimes described to be general-purpose (e.g. Radford et al., 2019), and are increasingly incorporated into real-world applications. However, their behavior can reflect a limited set of human knowledge and perspectives (Johnson et al., 2022; Durmus et al., 2023; Atari et al., 2023). Since the composition of pretraining data has been shown to impact model behavior (Kandpal et al., 2023; Razeghi et al., 2022; Chang et al., 2023; Gonen et al., 2023), documentation of this data facilitates informed and appropriate application of models (Gebru et al., 2021).

In our work, we argue that it is additionally important to examine how data is transformed prior to pretraining, and document the implications of these transformation steps. LLM pretraining data curation involves many decision points, which may be motivated by performance on popular benchmarks (e.g. Rae et al., 2021), or simply by some notion of text "quality." There remain many under-examined assumptions within data curation pipelines, which vary subtly across models (Soldaini et al., 2024; Penedo et al., 2023).

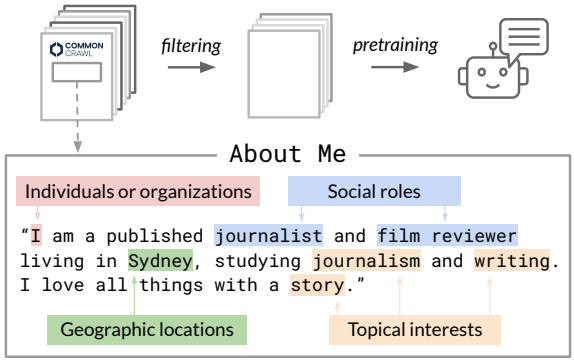

Figure 1: A paraphrased excerpt from a website's ABOUT page, with extracted social dimensions highlighted. We use self-descriptions like this one from Common Crawl, which is frequently used as LLM pretraining data, to examine the social effects of data curation filters.

We provide a new dataset and framework, AboutMe, for documenting data filtering's effects on web text grounded in social and geographic contexts. Sociolinguistic analyses in NLP are limited by a lack of large-scale, self-reported sociodemographic information tied to language data (Holstein et al., 2019; Andrus et al., 2021). Though text can be attributed to broad sources, e.g. Wikipedia, the backgrounds of content creators at more granular levels are often unknown. In particular, web crawl data lacks consistent and substantive user metadata. Our study leverages existing structure found in web data. Specifically, some websites include pages delineated to be *about* the website creator, such as an "about me" page (Figure 1). Thus, we are able

Statistic	Count
# of hostnames (websites)	10.3M
# of white-spaced tokens (ABOUT pages)	3.1B
# of white-spaced tokens (sampled pages)	3.5B
# of organizations	7.7M
# of individuals	2.6M
↳ # of individuals with labeled social roles	2.0M
# of hostnames labeled with country	6.5M

Table 1: A summary of count statistics for AboutMe.

to identify whose language is represented in web scraped text at an unprecedented scale.

From websites' ABOUT pages, we measure their topical interests, their positioning as individuals or organizations, their self-identified social roles, and their associated geographic locations (§2). We then apply ten "quality" and English ID filters drawn from prior literature on LLM development (§3) onto these websites, show whose pages are removed or retained, and investigate possible reasons for filters' preferences (§4). Together, our experiments uncover behavioral patterns within and across filters tied to aspects of websites' provenance. We find that model-based "quality" filters' implicit preferences for certain topical domains lead to text specific to different roles and occupations being removed at varying rates. In addition, English content associated with non-anglophone regions of the world can be removed due to filtering approaches that assume pages are monolingual.

We release our dataset, reproduced filters, and other resources to facilitate future work:

Code github.com/lucy3/whos_filtered
Dataset huggingface.co/datasets/allenai/aboutme

2 Extracting Social Dimensions from ABOUT Pages

Sociolinguists conceptualize language as a performance of one's *social identity*, or membership in a social group (Nguyen et al., 2016). Websites' ABOUT pages capture aspects of their creators' social identities that they deem salient and significant enough to mention in a summary (Figure 1). Thus, these self-descriptions can help delineate meaningful differences in language varieties and use. We extract social aspects that are present across large sets of pages using automated methods, some of which we contribute as novel approaches. We do not examine attributes such as race or gender, as these are less commonly explicitly stated on ABOUT pages and may raise the risk of mismeasurement (§8).

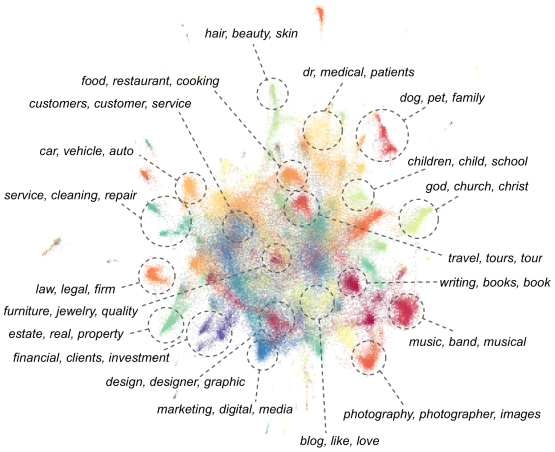

Figure 2: Examples of ABOUT web pages' topical interests annotated with cluster centers' top three representative words, obtained using an inverse transformation of cluster centroids and overlaid on a UMAP of pages. Appendix C lists all 50 topical clusters.

2.1 Data preprocessing

AboutMe is derived from twenty four public snapshots of Common Crawl collected between 2020–05 and 2023–06. We extract text using CCNet (Wenzek et al., 2020) and deduplicate URLs across all snapshots. Our study focuses on data curation of English LLMs, and our pipeline for identifying social aspects of websites uses methods that work best for English. Thus, we limit our study to CCNet's outputted webpages that have a fastText English ID score > 0.5 (Joulin et al., 2016b,a).

From this Common Crawl data, we identify websites that include an ABOUT page, or URL paths containing *about*, *about-me*, *about-us*, or *bio* (Appendix A). We then pair each ABOUT page with a random page on the same website. AboutMe thus contains both information about the creator/s of a website and a sample of their textual content (Table 1). Though we use this dataset to study LLM data curation practices, text linked to their creators' self-descriptions can also facilitate research on self-presentation (Sun et al., 2023; Pathak et al., 2021) and language variation (Nguyen et al., 2016).

2.2 Topical interests

First, we treat ABOUT pages as summaries of website creators' interests and topical focus. Following past work on the unsupervised discovery of domains (Gururangan et al., 2023), we embed ABOUT pages using unigram counts and tf-idf (Manning et al., 2008) and cluster them with balanced k-means. We set $k = 50$ and surface a wide range

of topical clusters in our data, including design, finance, food, religion, and travel (Figure 2, Appendix C). Thus, this clustering step provides an initial broad overview of who and what is in AboutMe.

2.3 Individuals vs. organizations

Website creators can range from casual bloggers to larger corporations. We deem webpages with urls that contain *about-me* or *bio* as individuals, while those that are labeled as *about-us* are organizations. However, some pages are ambiguously labeled with *about*, and so we classify these into individuals or organizations by training a binary random forest classifier on labeled pages. Classifier inputs include several count features that are agnostic to pages' topical content: the proportion of words on an ABOUT page in common pronoun series (*he*, *she*, *they*, *we*, *I*), the proportion of words that are tagged as a PERSON by spaCy named-entity recognition, and the number of unique PERSON first tokens. Our classifier achieves an average macro F1 of 89.2, via 5-fold cross validation on 10k examples per class. Hyperparameter choices, classifier confidence, and other implementation details are in Appendix D.1. By applying this classifier, we find that three-fourths of hostnames are organizations rather than individuals (Table 1).

2.4 Social roles

Among individuals, we extract their self-identified social roles from their ABOUT pages. The salience of a social role or occupation in a setting impacts language. Roles not only shift text's topical focus, but also facilitate the use of situation-specific language styles and registers (Agha, 2005).

String-matching can be imprecise due to polysemy and mentions of other people on ABOUT pages (e.g. a *customer*). Thus, our role extraction approach targets explicit expressions of self-identification (e.g. *I am a designer, entrepreneur, and mother*). We hand-label a sample of 1K ABOUT page sentences spanning a diverse set of potential roles,[1] and treat role extraction as a binary, sentence-level token classification task. Our full criteria for role annotation can be found in Appendix F.1, and we achieved high agreement (Cohen's $\kappa = 0.836$).

We finetune RoBERTa-base on our labeled data, as it provides a scalable yet flexible approach for learning a variety of self-identification patterns. Before finetuning, we continue pretraining RoBERTa on individuals' ABOUT pages, improving in-domain performance (Gururangan et al., 2020). Hyperparameters and model selection details can be found in Appendix F.2. Our best model achieves a F1 of 0.898 when evaluated at the word-level on a held-out test set.

With this approach, we are able to identify social roles on 77.7% of all individuals' ABOUT pages (Table 1), and pages that have any roles contain 5.5 on average (SD = 9.9). When possible, we group terms into occupations based on a taxonomy created by the Occupational Information Network, or O*NET (Peterson et al., 2001), e.g. the roles *attorney* and *lawyer* are in the occupation of *Lawyer* in the *Legal* occupation family (Table 2, Appendix F.3). For our filtering rate analysis (§4), we include 780 social roles that occur at least 1K times in AboutMe.

Occupation family	Count	Examples of extracted roles
Arts, Design, Entertainment, Sports, & Media	1.1M	*artist*, *director*, *designer*, *writer*, *photographer*, *musician*, *player*
Production	620K	*designer*, *engineer*, *maker*, *builder*, *operator*, *mechanic*
Community & Social Service	452K	*therapist*, *educator*, *advisor*, *pastor*, *activist*, *social worker*
Computer & Mathematical	365K	*engineer*, *developer*, *scientist*, *strategist*, *programmer*
Educational Instruction & Library	308K	*teacher*, *professor*, *lecturer*, *curator*, *tutor*, *graduate student*

Table 2: Five most common occupation families in AboutMe, by website count, with example social roles. Additional examples can be found in Appendix F.2-F.3.

2.5 Geography

Models' emphasis on English already restricts their ability to capture perspectives from people who write in other languages, especially populations outside of Western, anglophone countries (Blasi et al., 2022; Durmus et al., 2023). Still, English is commonly chosen for intercultural communication and sometimes characterized as a *world language* or *lingua franca* (Seidlhofer, 2005). Thus, our web-derived dataset includes a range of geographic contexts in which English is used.

We geoparse locations on ABOUT pages using Mordecai3, which tags named locations, retrieves candidate matches from a GeoNames index, and disambiguates them using textual context (Halterman, 2023). For example, *I'm from Alexandria, Virginia* would be geoparsed to a location in the United States instead of Egypt. Mordecai3 is free and uses a local index, and so it can be scaled

[1] https://en.wiktionary.org/w/index.php?title=Category:en:People

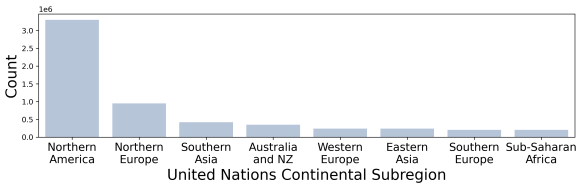

Figure 3: Common continental subregions in AboutMe. The most frequent countries are the United States, United Kingdom, India, Canada, Australia, China, Germany, New Zealand, Italy, and South Africa (Appendix G).

up to millions of webpages. With this approach, we are able to tag 63.2% of websites in AboutMe with at least one country. This coverage exceeds the 10.38% obtained by using country-specific top-level domains, e.g. .uk for the United Kingdom (Cook and Brinton, 2017).

Locations mentioned on a page are often associated in some way with the creator/s of a website, but the strength of this association can vary. For example, a company may have been founded in one country, but ships products to another. Through manual annotation of locations in 200 ABOUT pages, we find that 79.46% of websites with geoparsed countries originate from or reside in the most frequently referenced country on their ABOUT pages (evaluation details in Appendix G.1). Thus, we label each website with its most frequent country, yielding an overall website-level labeling accuracy of 91.0%. Due to the current global digital divide and our focus on English, the majority of websites in AboutMe are labeled with the United States and United Kingdom, with a long tail of other countries (Figure 3). For analysis, we group countries into 5 continental regions and 15 subregions delineated by the United Nations (Appendix G.2).

2.6 Summary

Overall, the websites in AboutMe cover a variety of topical interests, though a large proportion are associated with locations in the United States. A majority of websites are by organizations rather than individuals, and among individuals, most websites are created by people with creative and media-related occupations. Finally, though our methods for characterizing ABOUT pages achieve good performance, they still have limitations and mismeasurement risks, which we discuss in §7 and §8.

Filter	Examples of prior use	Removal strategy
⋆**WikiWebBooks**, or Wikipedia, OpenWebText, & Books3 classifier	GPT-3 (Brown et al., 2020)	Sampling based on scores
⋆**OpenWeb**, or Reddit outlinks classifier	the Pile (Gao et al., 2020)	Sampling based on scores
⋆**WikiRefs**, or Wikipedia references classifier	LLaMA (Touvron et al., 2023a) & RedPajama (Computer, 2023)	Cutoff: 0.25 (RedPajama), binary (LLaMA)
⋆**Wiki**, or Wikipedia classifier	Specified in reference mixes by Xie et al. (2023), PaLM (Chowdhery et al., 2023), and GPT-3 (Brown et al., 2020)	Sampling based on scores
⋆**Wiki**$_{ppl}$, or Wikipedia perplexity	CCNet (Wenzek et al., 2020)	Percentile cutoffs: 33.3% or 66.7%
⋆**Gopher** length, wordlist, repetition, & symbol rules	Gopher (Rae et al., 2021), Chinchilla (Hoffmann et al., 2022), & RefinedWeb (Penedo et al., 2023)	Specific cutoffs for each rule
∗**fastText** classifier	CCNet (Wenzek et al., 2020), LLaMA (Touvron et al., 2023a), RefinedWeb (Penedo et al., 2023)	Cutoffs: 0.50 (CCNet, LLaMA), 0.65 (RefinedWeb)
∗**CLD2** classifier	The Pile (Gao et al., 2020)	Cutoff: 0.50
∗**CLD3** classifier	multilingual C4 (Xue et al., 2021)	Cutoff: 0.70
∗**langdetect** classifier	C4 (Dodge et al., 2021; Raffel et al., 2023)	Cutoff: 0.99

Table 3: Quality filters (⋆) and langID systems (∗) investigated in our study.

3 Pretraining Data Filters

Raw data is transformed into pretraining data for LLMs through a variety of steps (Penedo et al., 2023; Soldaini et al., 2024, *inter alia*), including deduplication (Lee et al., 2022), decontamination (e.g. Touvron et al., 2023b), and explicit content filtering (e.g. OpenAI, 2023). We focus on analyzing the effects of "quality" filtering and English langID. The former is motivated by the subjectivity of how quality should be defined, and the latter by ongoing uncertainty around whether langID is robust to a wide range of language varieties (Seargeant and Tagg, 2011; Caswell et al., 2020).

Since web text contains noisy content (Eisenstein, 2013), removing or downsampling "low quality" text is common in LLM development (Table 3). However, mismatch between filtering outcomes and downstream objectives can lead to performance degradation on some tasks (Gao, 2021; Longpre et al., 2023), or disfavor content written by minoritized populations (Gururangan et al., 2022). LangID is also a common step in data curation pipelines (Table 3). It can be used at the document-level as an initial filter for language-specific (e.g. English-only) models, or to measure and adjust the composition of pretraining data for multilingual models (Xue et al., 2021). However, popular

langID systems are imperfect, for reasons such as training and application domain mismatch and confusion between similar languages (Kreutzer et al., 2022; Caswell et al., 2020).

We critically examine ten document-level quality and English filters that are sufficiently documented in prior work (Table 3). Appendix B includes additional details on the reproduction of each filter. Descriptions of pretraining data curation are sometimes too vague or non-existent to allow for exact replication (OpenAI, 2023), but multiple recent and prominent LLMs still allude to the use of model- and heuristic-based data filters (Touvron et al., 2023a; Gemini Team et al., 2023; Chowdhery et al., 2023).

Model-based quality. We experiment with quality filters that score text based on their similarity to some chosen "high quality" reference corpora. We name these filters based on the reference corpora used to train them: WIKIWEBBOOKS, OPENWEB, WIKI, and WIKIREFS (Table 3). We use Gururangan et al. (2022)'s replication of GPT-3's binary logistic regression quality classifier and only vary the positive "high quality" class. The negative class is a fixed set of tokens from the September 2019 dump of Common Crawl, and each class contains approximately 300M tokens. We also compare WIKI to a perplexity-based text scorer, WIKI_{ppl}, which uses a 5-gram Kneser-Ney language model trained on Wikipedia instead of a classifier (Wenzek et al., 2020; Laurençon et al., 2022; Muennighoff et al., 2023; Marion et al., 2023).

Heuristic-based quality. Another quality filtering approach for web text applies rule-based heuristics (Raffel et al., 2023; Rae et al., 2021). We examine 19 document-level heuristics and thresholds from Gopher (Raffel et al., 2023). These heuristics remove documents that do not meet thresholds pertaining to document and word length, textual repetition, and frequencies of symbols and common English words (Appendix B).

English langID. Our data is pre-filtered to documents that fastText langID scores as likely English (Joulin et al., 2016b,a). We also investigate whether the range of scores we observe with fastText generalize to other model-based measurements of English used in LLM development (Table 3). These langID systems include Compact Language Detector 2 (CLD2) (Sites, 2013), CLD3 (Salcianu et al., 2020), and langdetect (Shuyo, 2014).

4 Whose websites are filtered?

In this section, we overview the effects of data filters on sampled webpages grouped by social dimensions identified from their ABOUT pages. Broadly, we examine the degree of consensus among filters when scoring pages, and identify themes that characterize their behavior. Within some dimensions, we also investigate whether filtering rates reflect systemic differences in power and status among social groups (Blank, 2013; Davis, 2018).

Past LLMs have chosen a range of score cutoffs and sampling mechanisms to reduce undesirable text (Table 3). With this variation in mind, we examine the outcome of model-based filters through the lens of two contrasting scenarios. First, whose pages are least affected, or retained, if we were to keep only the documents within a top percentile of scores? Second, whose pages are most affected, or removed, if we were to filter those at a very bottom percentile? We select top and bottom cutoff percentiles of 10% and 90%, though for CLD2 and langdetect, a large number of score ties meant that that cutoffs for both scenarios were 5.2% and 8.7%, respectively. For rule-based filters, we use cutoffs specified by Rae et al. (2021). All cutoffs are listed in Appendix B.

4.1 Topical interests

Similarities in how data filters score topical clusters cut across quality and English filters (Table 4, Table 5). Pairwise correlations of topics' average English scores across all four langID systems have high consensus (mean r_s = 0.874, SD = 0.038, all $p < 0.001$). Surprisingly, Wikipedia perplexity also behaves like fastText langID (r_s = 0.860, $p < 0.001$). We qualitatively examine 20 random pages from highly filtered clusters, e.g. *fashion, women* and *online, store*. We find that pages with lower English scores list product names or specifications of individual products, and their original content may have been highly visual.[2] Indeed, further down the list of commonly highly filtered topical interests are clusters related to photography and art (Appendix C.2). Thus, though text-based LLMs may intend to be comprehensive in knowledge, they exclude information that is primarily communicated via other forms of media.

Differences among topical preferences show that the method of filtering and the choice of reference

[2]Indeed, manual inspection of current versions of these websites, when available, supports this claim.

Topical interests				Social roles				Geography			
least	**– rate**	**most**	**– rate**	**least**	**– rate**	**most**	**– rate**	**least**	**– rate**	**most**	**– rate**
law, legal	0.19	fashion, women	0.47	counsellor	0.16	jewelry designer	0.42	Northern Europe	0.26	Eastern Asia	0.31
blog, like	0.19	furniture, jewelry	0.42	hypnotherapist	0.16	production designer	0.40	Central Asia	0.26	Southern Asia	0.30
insurance, care	0.20	online, store	0.40	atheist	0.16	retoucher	0.40	Western Europe	0.26	South-eastern Asia	0.29
financial, clients	0.20	com, www	0.39	executive coach	0.17	illustrator	0.38	Northern America	0.26	Northern Africa	0.29
solutions, technology	0.20	products, quality	0.37	psychotherapist	0.17	concept artist	0.38	Australia & NZ	0.27	Western Asia	0.29

Table 4: The topical clusters, social roles, and geographic subregions that are least and most filtered by GOPHER heuristics. Appendix B.1 describes how individual rules affect webpages.

Quality: WIKIWEBBOOKS				Quality: OPENWEB				Quality: WIKIREFS			
↑ retained	**+ rate**	**↓ removed**	**– rate**	**↑ retained**	**+ rate**	**↓ removed**	**– rate**	**↑ retained**	**+ rate**	**↓ removed**	**– rate**
news, media	0.27	home, homes	0.21	news, media	0.32	estate, real	0.20	news, media	0.28	blog, like	0.21
film, production	0.24	estate, real	0.18	writing, books	0.20	home, homes	0.18	club, members	0.23	furniture, jewelry	0.20
writing, books	0.24	service, cleaning	0.18	software, data	0.20	furniture, jewelry	0.17	music, band	0.23	home, homes	0.19
research, university	0.22	blog, like	0.16	like, love	0.18	fashion, women	0.17	film, production	0.23	fashion, women	0.19
music, band	0.21	insurance, care	0.16	site, information	0.18	blog, like	0.16	research, university	0.22	service, cleaning	0.18
Quality: WIKI				**Quality: WIKI$_{ppl}$**				**English: fastText**			
↑ retained	**+ rate**	**↓ removed**	**– rate**	**↑ retained**	**+ rate**	**↓ removed**	**– rate**	**↑ retained**	**+ rate**	**↓ removed**	**– rate**
research, university	0.26	service, cleaning	0.22	law, legal	0.24	fashion, women	0.24	blog, like	0.22	fashion, women	0.21
film, production	0.25	home, homes	0.20	research, university	0.20	online, store	0.23	writing, books	0.22	online, store	0.20
music, band	0.21	insurance, care	0.16	god, church	0.19	quality, equipment	0.21	god, church	0.21	quality, equipment	0.18
art, gallery	0.21	marketing, digital	0.16	music, band	0.18	products, quality	0.21	photography, photographer	0.19	products, quality	0.18
law, legal	0.18	event, events	0.15	film, production	0.17	furniture, jewelry	0.20	like, love	0.19	furniture, jewelry	0.17
English: CLD2				**English: CLD3**				**English: langdetect**			
↑ retained	**+ rate**	**↓ removed**	**– rate**	**↑ retained**	**+ rate**	**↓ removed**	**– rate**	**↑ retained**	**+ rate**	**↓ removed**	**– rate**
insurance, care	0.97	quality, equipment	0.13	service, cleaning	0.22	fashion, women	0.19	blog, like	0.94	online, store	0.11
service, cleaning	0.97	company, products	0.09	life, yoga	0.19	quality, equipment	0.17	writing, books	0.93	fashion, women	0.11
law, legal	0.97	energy, water	0.09	like, love	0.18	online, store	0.17	life, yoga	0.93	quality, equipment	0.11
financial, clients	0.97	com, www	0.09	blog, like	0.18	art, gallery	0.16	god, church	0.93	products, quality	0.11
home, homes	0.97	research, university	0.08	dog, pet	0.17	products, quality	0.15	law, legal	0.93	com, www	0.11

Table 5: The result of simulating two contrasting filtering scenarios: which topical interests are *most retained* when all pages except those with the highest scores are filtered (*↑ retained*), and which are *most removed* when pages with the lowest scores are filtered (*↓ removed*). Numeric columns are topics' page removal (–) or retained rate (+). A few topical interests that recur throughout the table are highlighted for clarity. See Appendix C.2 for an extended and more detailed version of this table.

corpora can influence what "quality" entails. Despite both being trained on Wikipedia, a perplexity-based filter behaves differently from a linear classifier ($r_s = 0.382$, $p < 0.01$). In addition, a quality classifier's behavior reflects the composition of its reference corpora. For example, classifiers trained to prefer web content outlinked from Reddit or Wikipedia, including WIKIWEBBOOKS, OPENWEB, and WIKIREFS, highly score news and media websites. In contrast, WIKI and WIKIWEBBOOKS tend to prefer topics well-represented on Wikipedia, such as entertainment and science (Mesgari et al., 2015). Thus, these "quality" filters may optimize for topical domain fit.

4.2 Individuals vs. organizations

Research has suggested that "non-standard," colloquial language can be considered less desirable (Blodgett et al., 2020; Eisenstein, 2013). So, we hypothesize that organizations' language may be considered higher quality and more "English" than that of individuals, as organizations may have more editorial resources to professionally create their content (Wagner, 2002).

Surprisingly, we find that across all quality and English langID filters, web content created by individuals is widely preferred (Appendix D.2). For example, GOPHER removes 25.9% of webpages by individuals, and 28.3% of those by organizations. One reason for this pattern is that topics that receive overall low scores by data filters are dominated by organizations, e.g. businesses in the clusters *products, quality* and *online, store*. However, even when topics are fixed, organizations are still more likely to be removed by nearly all filters, with only WIKIREFS as an exception (Appendix D.2). Webpages by organizations tend to be shorter than those by individuals across and within topics, and their webpages include more non-alphabetic "words", more repetition, and fewer words from GOPHER's required list (all $p < 0.001$, Mann-Whitney U-test).[3]

[3] GOPHER's required wordlist includes *the*, *be*, *to*, *of*, *and*, *that*, *have*, *with*. See Appendix B for details.

Quality: WikiWebBooks				Quality: OpenWeb				Quality: WikiRefs			
↑ retained	+ rate	↓ removed	− rate	↑ retained	+ rate	↓ removed	− rate	↑ retained	+ rate	↓ removed	− rate
correspondent	0.38	home inspector	0.33	game developer	0.43	home inspector	0.31	correspondent	0.32	quilter	0.25
game developer	0.37	realtor	0.24	game designer	0.39	residential specialist	0.27	mayor	0.30	home inspector	0.24
game designer	0.36	real estate agent	0.23	data scientist	0.35	realtor	0.26	co-writer	0.30	crafter	0.24
essayist	0.34	inspector	0.23	correspondent	0.32	real estate broker	0.25	historian	0.30	stager	0.22
historian	0.34	stager	0.21	software engineer	0.34	real estate agent	0.25	bandleader	0.30	jewelry designer	0.21
Quality: Wiki				**Quality: Wiki$_{ppl}$**				**English: fastText**			
↑ retained	+ rate	↓ removed	− rate	↑ retained	+ rate	↓ removed	− rate	↑ retained	+ rate	↓ removed	− rate
laureate	0.35	wedding planner	0.21	law clerk	0.30	jewelry designer	0.17	christian	0.32	lighting designer	0.19
soprano	0.33	home inspector	0.20	litigator	0.26	lighting designer	0.16	catholic	0.31	production designer	0.18
conductor	0.32	momma	0.20	vice-chair	0.25	fashion designer	0.15	missionary	0.31	cinematographer	0.16
composer	0.31	dental assistant	0.20	conductor	0.24	production designer	0.14	mummy	0.29	retoucher	0.15
artistic director	0.30	mama	0.19	deputy	0.24	cinematographer	0.14	youth pastor	0.29	jewelry designer	0.15
English: CLD2				**English: CLD3**				**English: langdetect**			
↑ retained	+ rate	↓ removed	− rate	↑ retained	+ rate	↓ removed	− rate	↑ retained	+ rate	↓ removed	− rate
content strategist	0.99	laureate	0.13	counsellor	0.30	lighting designer	0.24	witch	0.96	production designer	0.11
home inspector	0.99	disciple	0.10	celebrant	0.28	production designer	0.23	barista	0.95	laureate	0.11
celebrant	0.99	soprano	0.10	hypnotherapist	0.25	sideman	0.21	naturopath	0.95	cinematographer	0.11
licensed professional counselor	0.98	language teacher	0.09	mummy	0.23	cinematographer	0.20	ally	0.95	retoucher	0.11
notary public	0.98	conductor	0.09	psychic	0.23	retoucher	0.19	cleaner	0.95	sideman	0.11

Occ. families: Arts, Design, Entertainment, Sports, & Media ; Community & Social Service ; Computer & Mathematical ; Sales & Related

Table 6: The result of simulating two contrasting filtering scenarios: which social roles are *most retained* when all pages except those with the highest scores are filtered (*↑ retained*), and which are *most removed* when pages with the lowest scores are filtered (*↓ removed*). Numeric columns include roles' page removal (−) or retained rate (+). For interpretation clarity, roles are highlighted if they belong to four frequently recurring O*NET occupation families. See Appendix F.4 for an extended and more detailed version of this table.

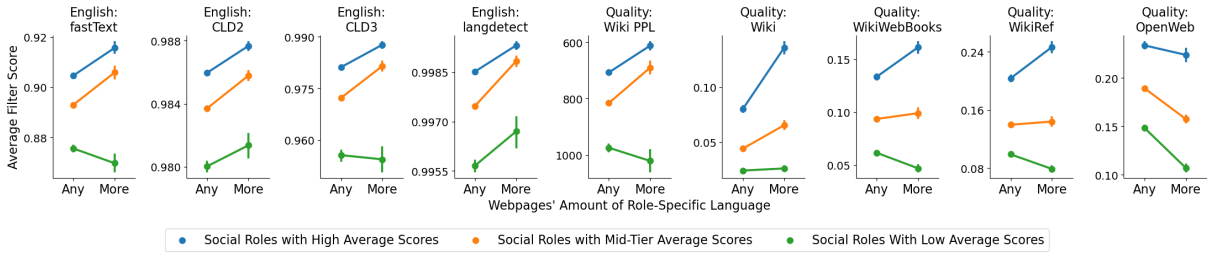

Figure 4: Webpages' use of role-specific words sometimes amplifies model-based filters' preferences. In each filter's plot, roles are bucketed into three tiers of high, mid, and low based on their overall average filter score, where higher values correspond to being less filtered. The first column in each plot is each tier's average filter score, while the second is after subsetting roles only to pages that use more role-specific words than average. Error bars are 95% CI over roles in each tier.

4.3 Social roles

Filters' social role preferences mirror those for topical interests (Table 6), such as programming occupations and *software, data* being preferred by OpenWeb, and *correspondent* matching the often "high quality" topic of *news, media*. We measure the relationship between model-based filter scores and two metrics that reflect occupations' societal status: their O*NET salary estimate and Hughes et al. (2022)'s survey-based ratings of prestige. We find small yet significant relationships between occupational prestige and model-based quality scores ($p < 0.001$, Appendix F.4). That is, pages linked to lower prestige occupations are filtered more by WikiWebBooks, OpenWeb, WikiRefs, Wiki, and Wiki$_{ppl}$.

The degree to which pages' self-identified roles are expressed through their text affects filtering as well (Figure 4). Within each role's collection of webpages, we calculate the proportion of each page that contains vocabulary specific to that role. Following past work (Zhang et al., 2017; Lucy et al., 2023), we identify role-specific vocabulary using a metric of association between word types and roles, where their normalized pointwise mutual information (NPMI) score is greater than 0.1. We compare how filters score all pages within a role, and how they score a subset of the role's pages that contain more role-specific words than average. We find that roles that are generally favored by a filter tend to be favored even more when their pages contain more role-specific words, in contrast to roles that are scored lowest by a filter, which do not benefit

from role-specific word use or are penalized further. Exceptions to this pattern are CLD2 and langdetect, two langID filters that score the vast majority (>90%) of pages similarly. These findings suggest that caution may be needed when using pretrained LLMs out-of-the-box for tasks and applications that involve language specific to some domains.

4.4 Geography

One striking commonality among several data filters is that they tend to assign low scores to webpages from Asia (Figure 5). For example, webpages are 2.4 times more likely to be removed by CLD2 if they are associated with Eastern Asia than Northern Europe. Eastern Asia is the most topically skewed subregion, as 29.2% of its websites are in the lowly-scored *quality, equipment* topic.

However, geographic filtering patterns are not only explained by topical differences. As expected, most English filters prefer subregions with "core anglophone" countries: Northern America (Canada and US), Northern Europe (UK), and Australia & New Zealand (Figure 5). Subregions with lower English document-level scores contain more non-English paragraphs across all langID systems (mean r = -0.791, SD = 0.084, all $p < 0.05$). By examining these "non-English" paragraphs, we observe two reasons for why a page may not be "English" enough. First, langID can mislabel English text (Caswell et al., 2020), such as content containing names of products, people, and non-anglophone locations. Second, some web pages are indeed multilingual, either code-switching or including multiple translations of the same content. LangID is usually applied at the document-level during data curation, and some systems may assume monolingual inputs (Zhang et al., 2018). Non-English paragraphs in `AboutMe` reflect their geography, e.g. Chinese in Eastern Asia, Spanish in Latin America, and Polish in Eastern Europe. Thus, simply choosing to communicate in English is not necessarily grounds for inclusion, and how English is situated within webpages matters.

In addition, we examine how filtering of geographic locations may relate to their relative global status. Past work has suggested that some NLP models may favor wealthier countries (Zhou et al., 2022). In our case, we do not find a significant relationship between a country's filter scores and their gross domestic product (Appendix G.3).

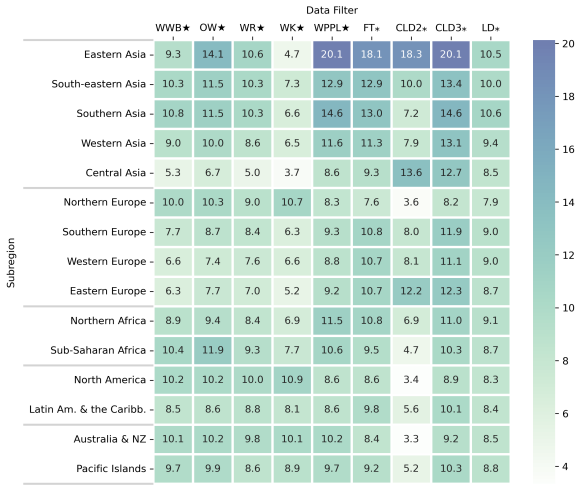

Figure 5: Webpage removal rates for each subregion when pages at a bottom percentile are removed by model-based filters, using cutoffs motivated in §4. Quality (⋆) and langID (⋆) filters in columns, left to right: WIKIWEBBOOKS, OPENWEB, WIKIREFS, WIKI, WIKI$_{ppl}$, fastText, CLD2, CLD3, and langdetect.

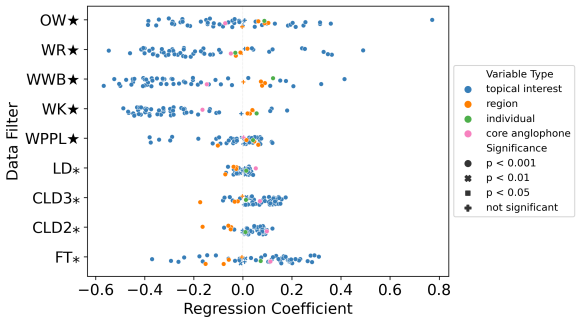

Figure 6: Coefficients for binary/categorical variables (x-axis) across nine regressions that predict webpages' quality (⋆) and English (⋆) scores (y-axis). More detailed numeric values can be found in Appendix H.

4.5 Regression analysis

Finally, we investigate the relative importance of the social dimensions highlighted in previous sections for model-based filters. That is, what matters more: who you are, or where you are from?

We run several ordinary least squares regressions with different filters' scores as dependent variables, and topical interests, continental regions, individual/organization status, core anglophone status, and pages' character length as independent variables. All independent variables, except for length, are binary or categorical. We z-score standardize each regression's dependent variable so that coefficients are similarly interpretable across them. We find that length has a positive effect ($p < 0.001$) on all model-based filter scores except for WIKI$_{ppl}$,

and many topical variables have stronger effects on filter scores than other variable types, especially among quality classifiers (Figure 6, Appendix H). Though earlier we noted that some subregions' filtering scores may be due to topical skew (§4.4), even when controlling for topic, Asia still has the most negative coefficients relative to other continental regions for all langID filters.

4.6 Summary

We find shared patterns in how quality and English filters score websites delineated by social aspects such as social roles and geography. Differences in how quality filters behave depend on both how they are implemented and their choice of "high quality" reference corpora (§4.1). This latter factor results in notions of "quality" being associated with certain topical domains e.g. news and media, and these topical preferences then lead to filters privileging content specific to some social roles over others' (§4.3). Finally, common langID classifiers can overlook English content in non-anglophone regions of the world, especially Asia, even when controlling for other variables such as webpage length and topic (§4.5).

5 Related Work

Our measurements of self-descriptions are most related to prior work studying online self-presentation. Online biographies are a particularly rich source of social identity markers. For example, Pathak et al. (2021) extracts personal identifiers, e.g. *farm wife* or *umass amherst '20*, from Twitter profile bios, and show that on aggregate, identifiers in these bios align with users' sociodemographic backgrounds. Others have extracted identifiers by splitting short form content by delimiters, e.g. *22yo | she/they* → {*22yo*, *she/they*} (Yoder et al., 2020; Pathak et al., 2021), or matching syntactic patterns, e.g. `person is X` (De-Arteaga et al., 2019; Madani et al., 2023). As one of the self-description analysis tools we employ, we contribute a novel and more flexible RoBERTa-based approach in §2.4 for extracting social role identifiers.

Though a nascent research area, analyses of pretraining data curation decisions have also been the focus of other recent literature (Longpre et al., 2023; Soldaini et al., 2024). For example, Gao (2021) showed that discarding too much pretraining data using the Pile's quality filter can lead to worse downstream task performance. Our work is closest in spirit to Gururangan et al. (2022), who use a dataset of high school newspapers to show that text from wealthier, more educated, and urban areas are more likely to be considered high quality by GPT-3's model-based quality filter. Similarly, concurrent work by Hong et al. (2024) found that image-text CLIP-filtering for visual language models excludes data from LGBTQ+ people, older women, and younger men at higher rates. We also critically examine social aspects of quality filtering at scale, but across a range of text filters.

6 Conclusion

In our work, we examine how ten "quality" and English langID filters used during LLM development affect web text created by a range of individuals and organizations with different topical interests, social roles, and geographic locations. To obtain this information, we use a new dataset of webpages that contain website creators' self-described social identities. Overall, our framework allows for model developers and practitioners to better understand whether and how their choice of filtering approach may affect the resulting composition of web data in unintended ways. Though some practices may seem tried and true for building powerful LLMs, we encourage future work to continue investigating, documenting, and mitigating their caveats and tradeoffs.

7 Limitations

Algorithmic measurements of websites allow our investigations to scale to millions of webpages. Still, we acknowledge that our dataset and analysis methods can also uphold language norms and standards that may disproportionately affect some social groups over others. For example, `AboutMe` consists of documents that meet a Fasttext langID English score threshold of 0.5, as the algorithmic tools we use for later analyses are created for English. There are likely some false negatives we excluded from analysis, as some English content may not meet this threshold. As another example, named entity recognition during geoparsing may rely on locations being stated using standard capitalization norms in text. Our study also focuses only on English, due to a current gap in multilingual tools for large-scale data documentation (Joshi et al., 2020). We hope that future work continues to improve these content analysis pipelines, especially for long-tail or minoritized language phenomena.

We study the effects of filters in isolation, but acknowledge that in practice, data curation steps are layered and combined. The exact preprocessing of text before filters are applied may impact outcomes; for example, some langID systems can be applied to web data prior to HTML removal (Gao et al., 2020). Unfortunately, commercially prominent LLMs often lack detailed documentation necessary for investigations at this level of specificity. Still, we encourage future work to investigate implications of layered LLM data curation practices.

8 Ethical Considerations

This work received IRB exemption, and includes several ethical considerations.

Measurement error. Our analysis approach leans more towards extraction of stated information and less towards inference of additional information. That is, we aim to minimize the extent to which we impose implicit labels on people. Still, we risk cases where websites are misidentified due to retrieval or identification error. Measurements of social identity from ABOUT pages are affected by reporting bias, where a lack of self-provided information can lead to pages being excluded from relevant analyses. We encourage future work to revisit these issues, while adhering to privacy-related principles in mind.

Pronouns. Exclusivity and misrepresentation harms towards non-binary people have been gaining attention in the NLP community (Dev et al., 2021; Cao and Daumé III, 2020). We recognize that in the process of measuring different aspects of *who* is filtered, websites by non-binary individuals are likely mishandled by the algorithmic approaches we use. That is, our classifier discerning individuals and organizations relies on common pronoun series as input features, but some non-binary people may use neopronouns, e.g. *xe/xem/xyr* (Lauscher et al., 2022; Ovalle et al., 2023). In addition, the models we leverage, such as spaCy and RoBERTa, may mishandle text containing neopronouns. Neopronouns, though rare, do appear in `AboutMe`; we surface approximately 21 websites whose ABOUT pages' most frequent pronoun series are neopronouns (Appendix E).

Intended use of dataset. Though the data we analyze is provided by Common Crawl, a source of publicly open web data, care still needs to be taken when handling this data. Future uses of this data should avoid incorporating personally identifiable information into generative models, report only aggregated results, and paraphrase quoted examples to protect the privacy of individuals (Bruckman, 2002).

Other considerations. Throughout this paper, we describe the effects of exclusion of data from pretraining as potentially perpetuating erasure or decreasing downstream model performance on relevant tasks. However, removal from pretraining data is not always a negative outcome. For example, in some cases, content creators may prefer that their content is not incorporated into training LLMs due to copyright violation and/or lack of consent.

9 Acknowledgements

We thank Nicholas Tomlin, Naitian Zhou, and Kyle Lo for helpful conversations and feedback. We also thank our anonymous reviewers for their thoughtful reviews. This work was supported in part by the National Science Foundation (IIS-1942591).

References

Asif Agha. 2005. *Registers of Language*, chapter 2. John Wiley & Sons, Ltd.

McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 249–260, New York, NY, USA. Association for Computing Machinery.

Mohammad Atari, Mona J Xue, Peter S Park, Damián E Blasi, and Joseph Henrich. 2023. Which humans?

Samuel Barham, Orion Weller, Michelle Yuan, Kenton Murray, Mahsa Yarmohammadi, Zhengping Jiang, Siddharth Vashishtha, Alexander Martin, Anqi Liu, Aaron Steven White, Jordan Boyd-Graber, and Benjamin Van Durme. 2023. Megawika: Millions of reports and their sources across 50 diverse languages.

Grant Blank. 2013. Who creates content? *Information, Communication & Society*, 16(4):590–612.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

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.

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.

Amy Bruckman. 2002. Studying the amateur artist: A perspective on disguising data collected in human subjects research on the internet. *Ethics and Information Technology*, 4:217–231.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, Singapore. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Together Computer. 2023. RedPajama: an open dataset for training large language models.

Paul Cook and Laurel J Brinton. 2017. Building and evaluating web corpora representing national varieties of English. *Language Resources and Evaluation*, 51:643–662.

Alexander Davis. 2018. *India and the Anglosphere: Race, identity and hierarchy in international relations*. Routledge.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia. Association for Computational Linguistics.

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

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv: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.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, ..., Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol

Vinyals. 2023. Gemini: A family of highly capable multimodal models.

Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in language models via perplexity estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148, Singapore. Association for Computational Linguistics.

H.P. Grice. 1975. Logic and conversation. *Syntax and Semantics*, 3:41–48.

Suchin Gururangan, Dallas Card, Sarah Dreier, Emily Gade, Leroy Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. 2022. Whose language counts as high quality? Measuring language ideologies in text data selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2562–2580, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Suchin Gururangan, Margaret Li, Mike Lewis, Weijia Shi, Tim Althoff, Noah A Smith, and Luke Zettlemoyer. 2023. Scaling expert language models with unsupervised domain discovery. *arXiv preprint arXiv:2303.14177*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Andrew Halterman. 2023. Mordecai 3: A neural geoparser and event geocoder. *arXiv preprint arXiv:2303.13675*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–16, New York, NY, USA. Association for Computing Machinery.

Rachel Hong, William Agnew, Tadayoshi Kohno, and Jamie Morgenstern. 2024. Who's in and who's out? a case study of multimodal clip-filtering in datacomp.

Bradley T Hughes, Sanjay Srivastava, Magdalena Leszko, and David M Condon. 2022. Occupational prestige: The status component of socioeconomic status.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an American accent: value conflict in GPT-3.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The BigScience ROOTS corpus: A 1.6 TB composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. 2023. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *arXiv preprint arXiv:2305.13169*.

Li Lucy, Jesse Dodge, David Bamman, and Katherine Keith. 2023. Words as gatekeepers: Measuring discipline-specific terms and meanings in scholarly publications. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6929–6947, Toronto, Canada. Association for Computational Linguistics.

Navid Madani, Rabiraj Bandyopadhyay, Briony Swire-Thompson, Michael Miller Yoder, and Kenneth Joseph. 2023. Measuring social dimensions of self-presentation in social media biographies with an identity-based approach.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, USA.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When less is more: Investigating data pruning for pretraining LLMs at scale. *arXiv preprint arXiv:2309.04564*.

Mostafa Mesgari, Chitu Okoli, Mohamad Mehdi, Finn Årup Nielsen, and Arto Lanamäki. 2015. "The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia. *Journal of the Association for Information Science and Technology*, 66(2):219–245.

Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling data-constrained language models. *arXiv preprint arXiv:2305.16264*.

Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska De Jong. 2016. Computational sociolinguistics: A survey. *Computational linguistics*, 42(3):537–593.

OpenAI. 2023. Gpt-4 technical report.

Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "I'm fully who I am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1246–1266, New York, NY, USA. Association for Computing Machinery.

Arjunil Pathak, Navid Madani, and Kenneth Joseph. 2021. A method to analyze multiple social identities in Twitter bios. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Norman G Peterson, Michael D Mumford, Walter C Borman, P Richard Jeanneret, Edwin A Fleishman, Kerry Y Levin, Michael A Campion, Melinda S Mayfield, Frederick P Morgeson, Kenneth Pearlman, et al. 2001. Understanding work using the occupational information network (o* net): Implications for practice and research. *Personnel Psychology*, 54(2):451–492.

Patrick Plonski, Asratie Teferra, and Rachel Brady. 2013. Why are more African countries adopting english as an official language. In *African Studies Association Annual Conference*, volume 23.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training Gopher. *arXiv preprint arXiv:2112.11446*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alex Salcianu, Andy Golding, Anton Bakalov, Chris Alberti, Daniel Andor, David Weiss, Emily Pitler, Greg

Coppola, Jason Riesa, Kuzman Ganchev, Michael Ringgaard, Nan Hua, Ryan McDonald, Slav Petrov, Stefan Istrate, and Terry Koo. 2020. Compact Language Detector v3 (CLD3). https://github.com/google/cld3?tab=readme-ov-file#credits.

Philip Seargeant and Caroline Tagg. 2011. English on the internet and a 'post-varieties' approach to language. *World Englishes*, 30(4):496–514.

Barbara Seidlhofer. 2005. English as a lingua franca. *ELT Journal*, 59(4):339–341.

Nakatani Shuyo. 2014. langdetect. https://github.com/Mimino666/langdetect.

Dick Sites. 2013. Compact Language Detector 2. https://github.com/CLD2Owners/cld2.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research.

Lu Sun, F Maxwell Harper, Chia-Jung Lee, Vanessa Murdock, and Barbara Poblete. 2023. Characterizing and identifying socially shared self-descriptions in product reviews. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 808–819.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Srdjan Vucetic. 2020. *The Anglosphere: A genealogy of a racialized identity in international relations*. Stanford University Press.

Ellen Wagner. 2002. Steps to creating a content strategy for your organization. *Best of the eLearning guild's learning solutions: Top articles from the eMagazine's first five years*, pages 103–120.

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 the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. 2023. Data selection for language models via importance resampling.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Michael Miller Yoder, Qinlan Shen, Yansen Wang, Alex Coda, Yunseok Jang, Yale Song, Kapil Thadani, and Carolyn P. Rosé. 2020. Phans, stans and cishets: Self-presentation effects on content propagation in tumblr. In *Proceedings of the 12th ACM Conference on Web Science*, WebSci '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Justine Zhang, William Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Community identity and user engagement in a multi-community landscape. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):377–386.

Yuan Zhang, Jason Riesa, Daniel Gillick, Anton Bakalov, Jason Baldridge, and David Weiss. 2018. A fast, compact, accurate model for language identification of codemixed text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 328–337, Brussels, Belgium. Association for Computational Linguistics.

Kaitlyn Zhou, Kawin Ethayarajh, and Dan Jurafsky. 2022. Richer countries and richer representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2074–2085, Dublin, Ireland. Association for Computational Linguistics.

A Data preprocessing

Our dataset, AboutMe, consists of ABOUT pages identified using webpage URLs (§2). Some webpages have multiple pages with URLs involving a target keyword (one of *about*, *about-me*, *about-us*, or *bio*). We retrieve ABOUT pages that end in /keyword/ or keyword.*, such as a URL ending in *about.html*. If there is only one of these candidates, we map the hostname to that one. If there are more than one, then we do not include that hostname in AboutMe, to avoid ambiguity around which page is actually about the main website creator. If a webpage has both *https* and *http* versions in Common Crawl, we take the *https* version.

Aside from cases where tokenizers are built into models or systems we use to analyze text, e.g. ROBERTA or Mordecai3, we use Microsoft's Bling Fire tokenizer.[4]

B Data filters

B.1 Filter reproduction

All model-based quality filters, except for WIKI$_{ppl}$, use the same implementation and parameter choices as the reproduction of GPT-3's quality filter by Gururangan et al. (2022).

WikiWebBooks. Both positive and negative examples for this classifier are the same as Gururangan et al. (2022). Their positive class consists of similar sized samples from Wikipedia, OpenWebText, and Books3. We reuse the same set of negative examples for other quality classifiers that share the same model architecture: OPENWEB, WIKIREFS, and WIKI.

OpenWeb. The original version of WebText was introduced in the GPT-2 paper, which described it as "all outbound links from Reddit, a social media platform, which received at least 3 karma" (Radford et al., 2019). We use an open and updated version of this dataset constructed by the Pile, called OpenWebText2 (Gao et al., 2020). The Pile uses this version to filter their version of Common Crawl. We sample documents from one shard of OpenWebText2 until we meet a 300M token ceiling to create the positive class for this classifier.

WikiRefs. We sample up to 300M tokens worth of webpages referenced by English Wikipedia to construct the positive class for this filter. We use text previously extracted by Barham et al. (2023).

[4]https://github.com/microsoft/BlingFire

Gopher heuristic	% of web pages affected
doclen	20.147
wordlen	0.942
symbol	0.135
bullet	0.039
ellipsis	1.083
alpha	3.529
stopword	9.723
repetition	13.361

Table 7: A breakdown of the effects of each Gopher rule on AboutMe's sampled webpages.

Wiki. We use text extracted from a dump of Wikipedia from March 20th, 2023.

Wiki$_{ppl}$. This perplexity-based KenLM filter trained on English Wikipedia is provided by CCNet, and its download link is specified in CCNet's Makefile.[5]

Gopher. We use Dolma's reproduction of Gopher's document-level rules for web text quality,[6] though we change median word length to mean word length to match the rule's description in the original Gopher paper (Rae et al., 2021). Table 7 overviews what percentages of webpages in AboutMe are removed by each rule or set of rules.[7] Overall, larger proportions of pages do not pass document length and repetition heuristics. Rules include the following, indicated by a shortened name for ease of reference:

- **doclen**: page length is between 50 and 100,000 words
- **wordlen**: mean word length is within 3 to 10 characters
- **symbol**: symbol-to-word ratio is less than 0.1, where symbols are either the hash symbol or ellipsis
- **bullet**: less than 90% of lines start with a bullet point
- **ellipsis**: less than 30% of lines end with an ellipsis
- **alpha**: more than 80% of words in a document contain at least one alphabetic character
- **stopword**: page contains at least two of the following English words: *the*, *be*, *to*, *of*, *and*, *that*, *have*, *with*

[5]https://github.com/facebookresearch/cc_net/blob/main/Makefile

[6]https://github.com/allenai/dolma

[7]Note that a single webpage can be affected by multiple rules.

Filter	↑ retained cutoff	↓ removed cutoff
fastText	≥ 0.97	< 0.68
CLD2	≥ 0.99	< 0.99
CLD3	≥ 1.0	< 0.9799
langdetect	≥ 1.0	< 1.0
WIKI$_{ppl}$	≥ 2225.7	< 268.1
WIKI	≥ 5.776e−2	< 1.298e−8
WIKIREFS	≥ 3.830e−1	< 2.422e−3
OPENWEB	≥ 4.307e−1	< 7.479e−3
WIKIWEBBOOKS	≥ 1.925e−1	< 8.981e−4

Table 8: Numerical cutoffs used for the two contrasting filtering scenarios motivated in §4.

- **repetition**: no content that exceeds several thresholds related to duplicated content: fraction of characters in most common bigrams (0.20), trigrams (0.18), or 4-grams (0.16), fraction of characters in duplicate 5-grams (0.15), 6-grams (0.14), 7-grams (0.13), 8-grams (0.12), 9-grams (0.11), 10-grams (0.10), fraction of duplicate lines (0.30), and fraction of characters in duplicate lines (0.20).

LangID. We build off of Dolma's toolkit to apply all langID filters to text (Soldaini et al., 2024). Dolma's existing functionality outputs English scores for CLD3, CLD2, and fasttext, and we implement analogous functions for applying langdetect. We also calculate paragraph- and sentence-level language scores for any language. For this, we follow Dolma's definition of a paragraph (character sequences separated by new lines) and sentence (Bling Fire's sentence tokenizer).

B.2 Score cutoffs

As described in the main text in §4, to investigate webpages that are most or least favored by a filter, we use two cutoffs: a more strict scenario that removes all but the top 10% of scores, and a more flexible one that removes only the bottom 10%. However, two filters, CLD2 and langdetect, contain many score ties, so we instead use the same numeric cutoff for both scenarios, and this cutoff corresponds to the bottom 5.2% and 8.7% percentiles of scores, respectively. Table 8 lists the numeric cutoffs that we used to obtain ↑ **retained** and ↓ **removed** results in the main paper (e.g. Table 6, Table 5, Figure 5). We observe that regression-based classifiers tend to label most Common Crawl webpages with low scores. In addition, among langID classifiers, fastText has the most graded and gradual score distribution, while other langID systems tend to mostly give very high or very low English scores.

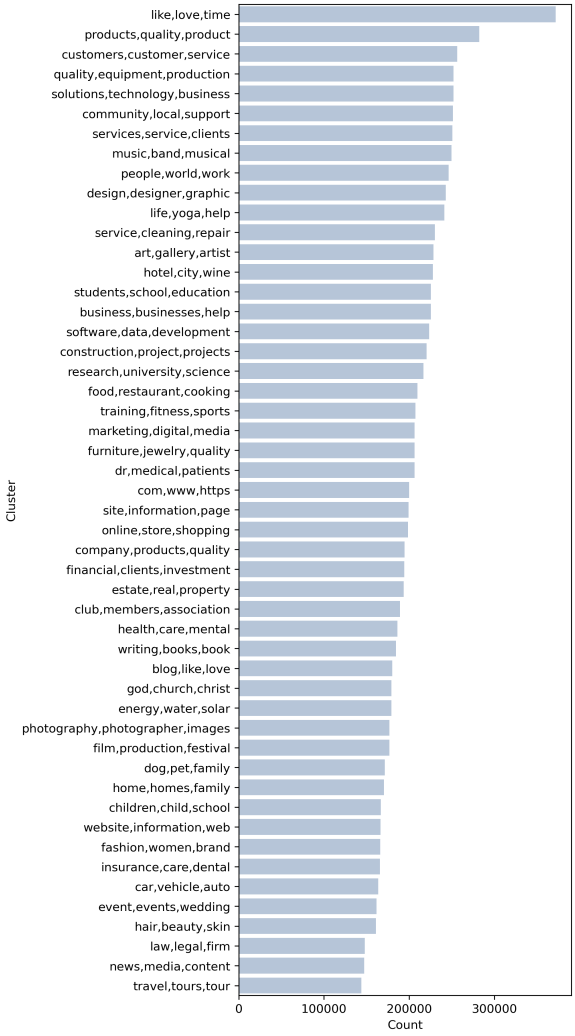

Figure 7: An ordered histogram of topical clusters' frequencies in `AboutMe`. Each topic is represented by their cluster centers' top three words.

C Topical interests

C.1 Clusters

For clustering, we use the same parameter choices as Gururangan et al. (2023).[8] We chose $k = 50$ as the number of clusters, because it offers a level of granularity that yields distinctive and interpretable topical areas. Since this version of k-means is balanced, clusters are encouraged to be similar in size. Figure 7 lists all 50 clusters and their frequency.

C.2 Additional filtering results

Table 9 shows the ten most and least Gopher-filtered topics, with a breakdown by rule. We find

[8] https://github.com/kernelmachine/cbtm

Most filtered topical interests			Least filtered topical interests		
Cluster	**- rate**	**Commonly "broken" rules**	**Cluster**	**- rate**	**Commonly "broken" rules**
fashion, women, brand	0.47	doclen (0.33), repetition (0.25), stopword (0.20)	law, legal, firm	0.19	doclen (0.13), repetition (0.09), stopword (0.05)
furniture, jewelry, quality	0.42	doclen (0.32), repetition (0.23), stopword (0.16)	blog, like, love	0.19	doclen (0.13), repetition (0.08), stopword (0.06)
online, store, shopping	0.40	doclen (0.26), repetition (0.24), stopword (0.17)	insurance, care, dental	0.20	doclen (0.15), repetition (0.09), stopword (0.06)
com, www, https	0.39	doclen (0.29), repetition (0.18), stopword (0.13)	financial, clients, investment	0.20	doclen (0.14), repetition (0.10), stopword (0.06)
products, quality, product	0.37	doclen (0.25), repetition (0.20), stopword (0.15)	solutions, technology, business	0.21	doclen (0.15), repetition (0.10), stopword (0.07)
art, gallery, artist	0.35	doclen (0.28), repetition (0.16), stopword (0.13)	dr, medical, patients	0.21	doclen (0.15), repetition (0.10), stopword (0.07)
photography, photographer, images	0.35	doclen (0.29), repetition (0.16), stopword (0.14)	health, care, mental	0.21	doclen (0.16), repetition (0.10), stopword (0.06)
customers, customer, service	0.33	doclen (0.23), repetition (0.17), stopword (0.13)	writing, books, book	0.21	doclen (0.16), repetition (0.10), stopword (0.06)
quality, equipment, production	0.32	doclen (0.21), repetition (0.17), stopword (0.14)	service, cleaning, repair	0.22	doclen (0.16), repetition (0.11), stopword (0.08)
food, restaurant, cooking	0.32	doclen (0.24), repetition (0.14), stopword (0.11)	travel, tours, tour	0.22	doclen (0.15), repetition (0.10), stopword (0.07)

Table 9: The top 10 most and least filtered topical interest clusters, with their removal rates, by Gopher heuristics. Numbers in parentheses indicate the fraction of documents in that topical cluster that are affected by a rule or set of rules, and the top three most common rules that affect pages in each topic are listed.

that the top three rules that affect pages within topics are similar; webpages from nearly all topics are highly filtered due to document length being too short. Table 10 is an extended version of Table 5, listing top ten topics instead of the top five.

D Individual and organizations

D.1 Classifier details

We separate out websites created by individuals and those by organizations using a random forest classifier. This classifier is trained on 10k randomly sampled *about me/bio* pages and 10k *about us* pages, and used to disambiguate *about* pages. It incorporates the following features:

- Proportion of words that are in each pronoun series: first person singular (*I*), first person plural (*we*), third person feminine (*she*), third person masculine (*he*), and third person gender-neutral/plural (*they*).
- Number of PERSON entities, normalized by the word length of the page
- Number of unique PERSON first tokens

To obtain named PERSON entities, we use spaCy's `en_core_web_trf` model. We set hyperparameters for our random forest classifier by selecting the best model based on its F1 score, cross validating over 5-folds, and performing randomized search over the following `scikit-learn` hyperpameters:

- `n_estimators`: 50, 100, 150, 200, 250, 300
- `criterion`: *entropy*, *gini*
- `max_depth`: None, 10, 50, 70, 100
- `min_samples_split`: 2, 5, 10, 20
- `min_samples_leaf`: 1, 2, 4.

Our best model with a F1 score of 0.892 had the following hyperparameters: `n_estimators` (200), `min_samples_split` (20), `min_samples_leaf` (2), `max_depth` (70), `criterion` (*gini*). Our resulting model tends to be highly confident based on its distribution of class probability scores (Figure 8). In other words, there are few websites that are around the border of what our model considers to be an organization or individual. Qualitatively, an example type of a website that is more ambiguous along the individual vs. organization dimension are businesses whose ABOUT pages tend to focus on the background of their current leader or founder.

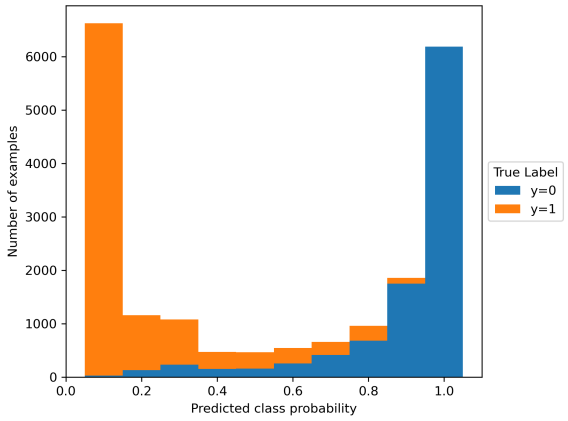

Figure 8: A stacked bar plot showing our individual and organization classifier's class probability scores across examples, colored by their true labels.

D.2 Additional filtering results

Table 11 shows filtering rates for individuals versus organizations for the two cutoff scenarios motivated in §4. On average, individuals have higher model-based scores than organizations for every filter, and so when the very top percentile of pages are retained, individuals are retained at higher rates. With regards to Gopher heuristics, 28.3% of organizations and 25.9% of individuals are removed, and the most prominent reason is again document

Quality: WikiWebBooks				Quality: OpenWeb				Quality: WikiRefs			
↑ retained	+ rate (% Δ)	↓ removed	− rate (% Δ)	↑ retained	+ rate (% Δ)	↓ removed	− rate (% Δ)	↑ retained	+ rate (% Δ)	↓ removed	− rate (% Δ)
news, media	0.27 (1.4→3.8)	home, homes	0.21 (1.7→1.5)	news, media	0.32 (1.4→4.5)	estate, real	0.20 (1.9→1.7)	news, media	0.28 (1.4→4.0)	blog, like	0.21 (1.7→1.5)
film, production	0.24 (1.7→4.2)	estate, real	0.18 (1.9→1.7)	writing, books	0.20 (1.8→3.6)	home, homes	0.18 (1.7→1.5)	club, members	0.23 (1.8→4.3)	furniture, jewelry	0.20 (2.0→1.8)
writing, books	0.24 (1.8→4.2)	service, cleaning	0.18 (2.2→2.0)	software, data	0.20 (2.2→4.3)	furniture, jewelry	0.17 (2.0→1.8)	music, band	0.23 (2.4→5.6)	home, homes	0.19 (1.7→1.5)
research, university	0.22 (2.1→4.7)	blog, like	0.16 (1.7→1.6)	like, love	0.18 (3.6→6.7)	fashion, women	0.17 (1.6→1.5)	film, production	0.23 (1.7→3.9)	fashion, women	0.19 (1.6→1.5)
music, band	0.21 (2.4→5.1)	insurance, care	0.16 (1.6→1.5)	site, information	0.18 (1.9→3.6)	blog, like	0.16 (1.7→1.6)	research, university	0.22 (2.1→4.7)	service, cleaning	0.18 (2.2→2.0)
club, members	0.17 (1.8→3.1)	furniture, jewelry	0.14 (2.0→1.9)	blog, like	0.18 (1.7→3.2)	quality, equipment	0.15 (2.4→2.3)	community, local	0.2 (2.4→4.8)	online, store	0.15 (1.9→1.8)
software, data	0.17 (2.2→3.6)	event, events	0.13 (1.6→1.5)	people, world	0.18 (2.4→4.3)	online, store	0.14 (1.9→1.8)	writing, books	0.18 (1.8→3.2)	hair, beauty	0.15 (1.6→1.5)
blog, like	0.16 (1.7→2.8)	fashion, women	0.12 (1.6→1.6)	film, production	0.16 (1.7→2.8)	products, quality	0.14 (2.7→2.6)	students, school	0.16 (2.2→3.6)	photography, photographer	0.15 (1.7→1.6)
site, information	0.16 (1.9→3.1)	construction, project	0.12 (2.1→2.1)	research, university	0.16 (2.1→3.4)	car, vehicle	0.13 (1.6→1.5)	site, information	0.14 (1.9→2.7)	products, quality	0.14 (2.7→2.6)
art, gallery	0.16 (2.2→3.6)	customers, customer	0.12 (2.5→2.4)	website, information	0.15 (1.6→2.4)	customers, customer	0.12 (2.5→2.4)	god, church	0.14 (1.7→2.4)	estate, real	0.14 (1.9→1.8)
Quality: Wiki				**Quality: Wiki$_{ppl}$**				**English: fastText**			
↑ retained	+ rate (% Δ)	↓ removed	− rate (% Δ)	↑ retained	+ rate (% Δ)	↓ removed	− rate (% Δ)	↑ retained	+ rate (% Δ)	↓ removed	− rate (% Δ)
research, university	0.26 (2.1→5.5)	service, cleaning	0.22 (2.2→1.9)	law, legal	0.24 (1.4→3.5)	fashion, women	0.24 (1.6→1.4)	blog, like	0.22 (1.7→3.8)	fashion, women	0.21 (1.6→1.4)
film, production	0.25 (1.7→4.2)	home, homes	0.2 (1.7→1.5)	research, university	0.20 (2.1→4.2)	online, store	0.23 (1.9→1.7)	writing, books	0.22 (1.8→3.8)	online, store	0.20 (1.9→1.7)
music, band	0.21 (2.4→5.2)	insurance, care	0.16 (1.6→1.5)	god, church	0.19 (1.7→3.3)	quality, equipment	0.21 (2.4→2.1)	god, church	0.21 (1.7→3.6)	quality, equipment	0.18 (2.4→2.2)
art, gallery	0.21 (2.2→4.6)	marketing, digital	0.16 (2.0→1.9)	music, band	0.18 (2.4→4.2)	products, quality	0.21 (2.7→2.4)	photography, photographer	0.19 (1.7→3.2)	products, quality	0.18 (2.7→2.5)
law, legal	0.18 (1.4→2.5)	event, events	0.15 (1.6→1.5)	film, production	0.17 (1.7→2.9)	furniture, jewelry	0.20 (2.0→1.8)	like, love	0.19 (3.6→6.6)	furniture, jewelry	0.17 (2.0→1.9)
club, members	0.17 (1.8→3.1)	car, vehicle	0.15 (1.6→1.5)	dr, medical	0.16 (2.0→3.1)	customers, customer	0.17 (2.5→2.3)	life, yoga	0.18 (2.3→4.2)	car, vehicle	0.16 (1.6→1.5)
news, media	0.17 (1.4→2.4)	business, businesses	0.14 (2.2→2.1)	community, local	0.16 (2.4→3.8)	company, products	0.14 (1.9→1.8)	dog, pet	0.17 (1.7→2.8)	customers, customer	0.15 (2.5→2.3)
writing, books	0.15 (1.8→2.7)	services, service	0.14 (2.4→2.3)	writing, books	0.15 (1.8→2.7)	car, vehicle	0.13 (1.6→1.5)	children, child	0.17 (1.6→2.6)	com, www	0.15 (1.9→1.8)
community, local	0.14 (2.4→3.5)	website, information	0.13 (1.6→1.6)	students, school	0.15 (2.2→3.2)	com, www	0.12 (1.9→1.9)	music, band	0.15 (2.4→3.6)	company, products	0.13 (1.9→1.8)
students, school	0.14 (2.2→3.1)	estate, real	0.13 (1.9→1.8)	financial, clients	0.15 (1.9→2.7)	hair, beauty	0.12 (1.6→1.5)	law, legal	0.15 (1.4→2.1)	art, gallery	0.12 (2.2→2.2)
English: CLD2				**English: CLD3**				**English: langdetect**			
↑ retained	+ rate (% Δ)	↓ removed	− rate (% Δ)	↑ retained	+ rate (% Δ)	↓ removed	− rate (% Δ)	↑ retained	+ rate (% Δ)	↓ removed	− rate (% Δ)
insurance, care	0.97 (1.6→1.7)	quality, equipment	0.13 (2.4→2.3)	service, cleaning	0.22 (2.2→4.3)	fashion, women	0.19 (1.6→1.5)	blog, like	0.94 (1.7→1.8)	online, store	0.11 (1.9→1.9)
service, cleaning	0.97 (2.2→2.3)	company, products	0.09 (1.9→1.8)	life, yoga	0.19 (2.3→3.9)	quality, equipment	0.17 (2.4→2.3)	writing, books	0.93 (1.8→1.8)	fashion, women	0.11 (1.6→1.6)
law, legal	0.97 (1.4→1.5)	energy, water	0.09 (1.7→1.7)	like, love	0.18 (3.6→5.6)	online, store	0.17 (1.9→1.8)	life, yoga	0.93 (2.3→2.4)	quality, equipment	0.11 (2.4→2.4)
financial, clients	0.97 (1.9→1.9)	com, www	0.09 (1.9→1.9)	blog, like	0.18 (1.7→2.7)	art, gallery	0.16 (2.2→2.1)	god, church	0.93 (1.7→1.8)	products, quality	0.11 (2.7→2.7)
home, homes	0.97 (1.7→1.7)	research, university	0.08 (2.1→2.0)	dog, pet	0.17 (1.7→2.5)	products, quality	0.15 (2.7→2.6)	law, legal	0.93 (1.4→1.5)	com, www	0.11 (1.9→1.9)
health, care	0.97 (1.8→1.8)	website, information	0.07 (1.6→1.6)	insurance, care	0.17 (1.6→2.4)	furniture, jewelry	0.15 (2.0→1.9)	health, care	0.93 (1.8→1.8)	furniture, jewelry	0.11 (2.0→2.0)
dog, pet	0.96 (1.7→1.7)	site, information	0.07 (1.9→1.9)	home, homes	0.17 (1.7→2.4)	music, band	0.14 (2.4→2.3)	like, love	0.93 (3.6→3.7)	customers, customer	0.1 (2.5→2.4)
life, yoga	0.96 (2.3→2.4)	online, store	0.07 (1.9→1.9)	site, information	0.17 (1.9→2.8)	photography, photographer	0.14 (1.7→1.6)	children, child	0.92 (1.6→1.6)	car, vehicle	0.1 (1.6→1.6)
god, church	0.96 (1.7→1.8)	art, gallery	0.07 (2.2→2.2)	law, legal	0.16 (1.4→2.0)	com, www	0.14 (1.9→1.9)	people, world	0.92 (2.4→2.4)	company, products	0.1 (1.9→1.9)
construction, project	0.96 (2.1→2.2)	fashion, women	0.07 (1.6→1.6)	website, information	0.16 (1.6→2.3)	film, production	0.14 (1.7→1.6)	financial, clients	0.92 (1.9→1.9)	energy, water	0.09 (1.7→1.7)

Table 10: The result of simulating two contrasting filtering scenarios for each filter (§4): which topical interests are *most retained* when all pages except those with the highest scores are filtered (↑ *retained*), and which are *most removed* when pages with the lowest scores are filtered (↓ *removed*). Numeric columns include topics' page removal rate (−) or retained rate (+), and their percentages in the dataset before and after applying a cutoff (% Δ). Topical interests that recur as the most or least preferred throughout the table are highlighted.

Quality: WikiWebBooks				Quality: OpenWeb				Quality: WikiRefs			
↑ retained	+ rate (% Δ)	↓ removed	− rate (% Δ)	↑ retained	+ rate (% Δ)	↓ removed	− rate (% Δ)	↑ retained	+ rate (% Δ)	↓ removed	− rate (% Δ)
individuals	0.15 (25.0→37.0)	organizations	0.1 (75.0→74.7)	individuals	0.14 (25.0→34.5)	individuals	0.1 (25.0→25.0)	individuals	0.11 (25.0→27.6)	individuals	0.11 (25.0→24.7)
organizations	0.08 (75.0→63.0)	individuals	0.09 (25.0→25.3)	organizations	0.09 (75.0→65.5)	organizations	0.1 (75.0→75.0)	organizations	0.1 (75.0→72.4)	organizations	0.1 (75.0→75.3)
Quality: Wiki				**Quality: Wiki$_{ppl}$**				**English: fastText**			
↑ retained	+ rate (% Δ)	↓ removed	− rate (% Δ)	↑ retained	+ rate (% Δ)	↓ removed	− rate (% Δ)	↑ retained	+ rate (% Δ)	↓ removed	− rate (% Δ)
individuals	0.12 (25.0→30.3)	organizations	0.11 (75.0→74.4)	individuals	0.12 (25.0→30.3)	organizations	0.11 (75.0→74.3)	individuals	0.17 (25.0→41.7)	organizations	0.1 (75.0→74.5)
organizations	0.09 (75.0→69.7)	individuals	0.08 (25.0→25.6)	organizations	0.09 (75.0→69.7)	individuals	0.08 (25.0→25.7)	organizations	0.08 (75.0→58.3)	individuals	0.08 (25.0→25.5)
English: CLD2				**English: CLD3**				**English: langdetect**			
↑ retained	+ rate (% Δ)	↓ removed	− rate (% Δ)	↑ retained	+ rate (% Δ)	↓ removed	− rate (% Δ)	↑ retained	+ rate (% Δ)	↓ removed	− rate (% Δ)
individuals	0.95 (25.0→25.2)	organizations	0.05 (75.0→74.8)	individuals	0.12 (25.0→26.4)	organizations	0.1 (75.0→74.8)	individuals	0.92 (25.0→25.3)	organizations	0.09 (75.0→74.7)
organizations	0.95 (75.0→74.8)	individuals	0.05 (25.0→25.2)	organizations	0.11 (75.0→73.6)	individuals	0.09 (25.0→25.2)	organizations	0.91 (75.0→74.7)	individuals	0.08 (25.0→25.3)

Table 11: The result of simulating two contrasting filtering scenarios for each filter (§4): who is *most retained* when all pages except those with the highest scores are filtered (↑ *retained*), and who are *most removed* when pages with the lowest scores are filtered (↓ *removed*). Numeric columns include individuals' or organizations' page removal rate (−) or retained rate (+), and their percentages in the dataset before and after applying a cutoff (% Δ).

Gopher heuristic	% of organizations	% of individuals
doclen	20.31	19.67
wordlen	0.98	0.84
symbol	0.11	0.20
bullet	0.04	0.03
ellipsis	0.99	1.36
alpha	3.69	3.04
stopword	10.08	8.64
repetition	14.06	11.27

Table 12: A breakdown of the effects of each Gopher rule on individuals and organizations.

length (Table 12). Across most filters, individuals within each topic have, on average, higher scores than organizations in the same topic (Table 13).

Filter	Fraction of topics	Majority?
fastText	0.84	✓
CLD2	0.64	✓
CLD3	0.68	✓
langdetect	0.60	✓
Wiki$_{ppl}$	0.94	✓
Wiki	0.60	✓
WikiRefs	0.32	×
OpenWeb	0.86	✓
WikiWebBooks	0.92	✓

Table 13: The percentage of topics where individuals have significantly higher model-based filter scores on average than organizations in the same topic (Mann Whitney U-test $p < 0.001$). Note that for Wiki$_{ppl}$, we reverse perplexity scores so that the higher, the better, to match the same direction as other model-based filters.

E Neopronouns

Our individual versus organization classifier uses pronoun counts as input features. During the process of gathering these pronoun features, we also examined possible ways to quantify or extract neopronouns from `AboutMe`. We began with an initial list of pronoun series that includes common neopronouns.[9] Some of these pronoun series, such as *it/it/its* and *kit/kit/kits* lead to a high number of false positives with exact string matching. In addition, we were not able to disambiguate cases where *they/them/theirs* is used as a plural pronoun instead of a singular one.

For other pronoun series, we identify potential pages whose subject uses neopronouns by finding ABOUT pages that include at least two unique pronouns from a neopronoun series, and that neopronoun series' frequency exceeds the frequency of more common pronouns. We manually inspected a sample of pages for each neopronoun series extracted with this approach, and estimate that only ~21 of 10.3M ABOUT pages contain uses of neopronoun terms as pronouns. The most frequent neopronoun series is *xe/xem/xyr*, with 8 extracted occurrences. Overall, the counts we obtained were too low for inclusion in our study. They are also likely an undercount, as we were only able to verify for precision rather than recall. We encourage future work to consider safe and inclusive studies of pronoun use in self-descriptions.

F Social roles

F.1 Annotation

We begin by annotating sentences that possibly contain terms referring to people. We explored two possible options for obtaining a seed list of terms: English Wiktionary's `Category:en:People`, and WordNet hyponyms of *person*. We found that the latter is imprecise (e.g. WordNet lists *have* as a *person* due to the phrase *haves and have-nots*) and outdated. So, we used English Wiktionary's list as a starting point for capturing a wide and up-to-date range of social roles (e.g. *influencer*). After removing terms that are overly long (4+-grams),[10] we string-matched for 10,676 Wiktionary terms on individuals' ABOUT pages. To avoid overfitting to popular roles, we reservoir sample for one ABOUT page per term, and then sampled 1000 random examples from that pool for annotation. We annotate head tokens of roles in the context of a single sentence, with seed terms pre-highlighted for annotators to verify, add to, or remove. We divide examples for annotation among the authors of this paper, following instructions shown in Figure 11.

Across all 1k annotated sentences, 541 contain at least positively labeled one social role in them. Overall, our annotators marked 1284 unique spans as roles. Thirty-five sentences were doubly annotated. Our annotators had good sentence-level agreement (Cohen $\kappa = 0.836$), and only differed on 4 of these sentences total.

F.2 Token classification

For finetuning RoBERTa, we grid-search through several learning rate options (1e-5, 2e-5, and 3e-5), experiment with varying levels of continued masked-language-modeling pretraining, and use a train-dev-test split of 600/200/200 labeled examples (Table 15). For other parameters, we use the same choices as Gururangan et al. (2020).

Our labeled spans are whole words, but RoBERTa sometimes labels parts of words. When we run inference on all individuals' ABOUT pages, we find that it nearly always tags all wordpieces in a positive span correctly. Still, 3.3% of tagged words are partially tagged, e.g. *play-**mate***, *trend-**set-ter***, ***mom**-my*. From manual inspection of these cases, it seems like partially tagged words are usually social roles. Thus, we evaluate at the word-level, and count words as social roles if any of its wordpieces is tagged as one.

Table 16 shows common terms extracted from all individuals' ABOUT pages. Some tagged words are part of hyphenated phrases, e.g. ***co**-president*. We do not consider common prefixes (e.g. *vice-*, *ex-*) and suffixes (e.g. *-elect*, *-in-law*) as individual roles during analysis.

F.3 Occupation hierarchy

Some of the roles we analyze are occupations, which we define as job titles grouped by the Occupational Information Network, or O*NET, which is created by the U.S. Department of Labor (Table 17). Job titles for occupations listed in O*NET are obtained from three sources. First, occupation pages themselves contain example job titles in singular form, usually in a comma-separated list. Second, the names of occupations often refer to job titles, e.g. ***Plasterers** and **Stucco Masons***, though in plu-

[9] https://github.com/witch-house/pronoun.is/blob/master/resources/pronouns.tab

[10] These tend to be sayings such as *life of the party* or *big fish in a small pond*.

Quality: WikiWebBooks				Quality: OpenWeb				Quality: WikiRefs			
↑ retained	+ rate (# docs)	↓ removed	− rate (# docs)	↑ retained	+ rate (# docs)	↓ removed	− rate (# docs)	↑ retained	+ rate (# docs)	↓ removed	− rate (# docs)
correspondent	0.38 (1438)	home inspector	0.33 (564)	game developer	0.43 (723)	home inspector	0.31 (527)	correspondent	0.32 (1213)	quilter	0.25 (322)
game developer	0.37 (618)	realtor	0.24 (7413)	game designer	0.39 (707)	residential specialist	0.27 (419)	mayor	0.30 (667)	home inspector	0.24 (412)
game designer	0.36 (653)	real estate agent	0.23 (4726)	data scientist	0.35 (952)	realtor	0.26 (8291)	co-writer	0.30 (337)	crafter	0.24 (732)
essayist	0.34 (353)	inspector	0.23 (870)	correspondent	0.32 (1197)	real estate broker	0.25 (2273)	historian	0.30 (2224)	stager	0.22 (263)
historian	0.34 (2492)	stager	0.21 (259)	software engineer	0.34 (10436)	real estate agent	0.25 (5004)	bandleader	0.30 (445)	jewelry designer	0.21 (280)
laureate	0.32 (461)	residential specialist	0.21 (330)	full stack developer	0.31 (401)	salesperson	0.24 (642)	co-producer	0.30 (454)	mommy	0.21 (754)
reporter	0.32 (3581)	real estate broker	0.21 (1878)	hacker	0.31 (503)	sales associate	0.23 (364)	sideman	0.30 (533)	newbie	0.2 (215)
atheist	0.32 (341)	sales associate	0.19 (303)	atheist	0.31 (325)	broker	0.23 (4599)	soprano	0.30 (891)	shopper	0.2 (264)
playwright	0.32 (1246)	broker	0.19 (3890)	coder	0.28 (555)	inspector	0.22 (838)	conductor	0.29 (1575)	handyman	0.19 (205)
co-writer	0.31 (349)	salesperson	0.19 (512)	reporter	0.28 (3084)	quilter	0.21 (274)	record producer	0.28 (328)	knitter	0.19 (305)
Quality: Wiki				**Quality: Wiki$_{ppl}$**				**English: fastText**			
↑ retained	+ rate (# docs)	↓ removed	− rate (# docs)	↑ retained	+ rate (# docs)	↓ removed	− rate (# docs)	↑ retained	+ rate (# docs)	↓ removed	− rate (# docs)
laureate	0.35 (493)	wedding planner	0.21 (400)	law clerk	0.30 (497)	jewelry designer	0.17 (226)	christian	0.32 (2644)	lighting designer	0.19 (212)
soprano	0.33 (1001)	home inspector	0.2 (336)	litigator	0.26 (437)	lighting designer	0.16 (183)	catholic	0.31 (369)	production designer	0.18 (195)
conductor	0.32 (1743)	momma	0.2 (409)	vice-chair	0.25 (338)	fashion designer	0.15 (919)	missionary	0.31 (680)	cinematographer	0.16 (728)
composer	0.31 (8429)	dental assistant	0.20 (210)	conductor	0.24 (1321)	production designer	0.14 (157)	mummy	0.29 (375)	retoucher	0.15 (171)
artistic director	0.3 (2397)	mama	0.19 (1581)	deputy	0.24 (270)	cinematographer	0.14 (638)	youth pastor	0.29 (295)	jewelry designer	0.15 (193)
production designer	0.29 (315)	mommy	0.19 (690)	arbitrator	0.24 (264)	retoucher	0.13 (154)	oldest	0.29 (471)	mixer	0.14 (179)
improviser	0.29 (417)	mummy	0.18 (233)	clinical professor	0.23 (294)	artisan	0.13 (235)	atheist	0.28 (296)	set designer	0.14 (261)
research fellow	0.28 (1359)	mortgage broker	0.18 (188)	attorney, lawyer	0.23 (8569)	concept artist	0.13 (181)	baby	0.28 (575)	soprano	0.14 (421)
co-writer	0.28 (308)	couple	0.17 (182)	clerk	0.23 (1055)	set designer	0.12 (221)	freshman	0.28 (582)	sideman	0.14 (247)
arranger	0.28 (1769)	gal	0.17 (740)	historian	0.23 (1665)	colorist	0.12 (183)	sister	0.27 (2201)	fashion designer	0.13 (834)
English: CLD2				**English: CLD3**				**English: langdetect**			
↑ retained	+ rate (# docs)	↓ removed	− rate (# docs)	↑ retained	+ rate (# docs)	↓ removed	− rate (# docs)	↑ retained	+ rate (# docs)	↓ removed	− rate (# docs)
content strategist	0.99 (1013)	laureate	0.13 (181)	counsellor	0.30 (3243)	lighting designer	0.24 (280)	witch	0.96 (1311)	production designer	0.11 (122)
home inspector	0.99 (1661)	disciple	0.10 (134)	celebrant	0.28 (571)	production designer	0.23 (252)	barista	0.95 (1121)	laureate	0.11 (154)
celebrant	0.99 (1982)	soprano	0.10 (289)	hypnotherapist	0.25 (1372)	sideman	0.21 (378)	naturopath	0.95 (1411)	cinematographer	0.11 (504)
licensed professional counselor	0.98 (3848)	language teacher	0.09 (93)	mummy	0.23 (300)	cinematographer	0.20 (932)	ally	0.95 (1307)	retoucher	0.11 (122)
notary public	0.98 (1091)	conductor	0.09 (488)	psychic	0.23 (404)	retoucher	0.19 (220)	cleaner	0.95 (1028)	sideman	0.11 (189)
licensed clinical social worker	0.98 (3204)	artistic director	0.09 (690)	psychotherapist	0.23 (2445)	set designer	0.19 (354)	beginner	0.95 (1276)	artisan	0.1 (183)
beauty therapist	0.98 (1295)	improviser	0.08 (123)	channel	0.22 (265)	soprano	0.19 (569)	youth worker	0.94 (1491)	design director	0.1 (139)
lcsw	0.98 (1585)	curator	0.08 (1043)	life coach	0.22 (3385)	saxophonist	0.18 (387)	youth	0.94 (956)	3d artist	0.1 (166)
mental health counselor	0.98 (2819)	grandson	0.08 (83)	family therapist	0.22 (1488)	laureate	0.18 (255)	private tutor	0.94 (2176)	photo editor	0.1 (127)
communications director	0.98 (1103)	translator	0.08 (796)	mum	0.22 (2671)	bandleader	0.18 (261)	feminist	0.94 (2180)	soprano	0.10 (300)

Occupation families, by color: Arts, Design, Entertainment, Sports, and Media; Community and Social Service; Computer and Mathematical; Sales and Related

Table 14: The result of simulating two contrasting filtering scenarios for each filter (§4): which roles/occupations are *most retained* when all pages except those with the highest scores are removed (↑ *retained*), and which are *most filtered* when pages with the lowest scores are removed (↓ *removed*). Numeric columns include roles/occupations' page removal rate (−) or retained rate (+), and the # of documents removed or retained in parentheses. For interpretation clarity, occupations are highlighted in color if they belong to four frequently recurring O*NET occupation families.

Learning rate	Further pretraining?	Precision	Recall	F1
1e-5	None	0.797	0.958	0.870
	1 epoch	0.814	0.958	0.880
	10 epoch	0.805	0.971	0.880
2e-5	None	0.792	0.941	0.860
	1 epoch	0.842	0.937	0.887
	10 epoch	0.835	0.958	0.892
3e-5	None	0.806	0.945	0.870
	1 epoch	0.827	0.924	0.873
	10 epoch	**0.856**	**0.945**	**0.898**

Table 15: Performance of RoBERTa-base models on a role classification task, with our chosen model's scores bolded.

ral. We singularize and parse these occupation names into job titles by querying GPT-3.5 with the prompt template shown in Figure 9. We manually verify answers from GPT-3.5 that do not agree with a simple rule-based approach of splitting occupations on commas and *and* and removing *-s* from terms in occupation titles. Finally, we obtain additional job titles for each O*NET occupation by including all job titles listed in O*NET's file of "alternate" or "lay" occupational titles.[11]

[11] https://www.onetcenter.org/dictionary/20.3/text/alternate_titles.html

```
Split the following list of occupations and
convert to singular: Watch and Clock Repairers

Answer: watch repairer, clock repairer

Split the following list of occupations and
convert to singular: Plumbers, Pipefitters, and
Steamfitters

Answer: plumber, pipefitter, steamfitter

Split the following list of occupations and
convert to singular: [insert occupation name
here]

Answer:
```

Figure 9: Prompt for reformatting occupation names into a series of job titles.

Since English tends to have head-final noun phrases, we attach ABOUT pages to job titles if their last token is classified as a role. Social roles have varying levels of granularity and one term can link to multiple, e.g a *floral designer* is both a *designer* due to its head token and a *florist* due to O*NET. Some commonly extracted social roles (e.g. *student*, *mom*) are not in O*NET, so we analyze scores for each of these terms individually. Other terms are ambiguous as to which O*NET

occupation they refer to (e.g. a *researcher* could be a historian or a geneticist), and so we analyze these individually as well.

F.4 Additional filtering results

Table 14 shows an extended version of Table 6.[12]

For our prestige and salary analyses, we gather salary estimates from O*NET occupation pages, and prestige ratings from Hughes et al. (2022). For ambiguous O*NET job titles (e.g. *researcher*), we assign them the average salary and prestige of all occupations they belong to. One limitation of this metadata is that these salary and prestige estimates are gathered from a U.S.-centric perspective, and may not generalize to other geographic contexts. Out of all 780 unique social roles that occur more than 1K times in AboutMe, 462 (59.2%) have prestige values and 497 (63.7%) have salaries.

We find that for salary, two filters (WIKI$_{ppl}$ and Gopher) show statistically significant relationships with salary, where higher-paid occupations are filtered less (Table 18, Figure 10). For prestige, all quality filters except for Gopher show a statistically significant relationship. The most and least filtered social roles shown in Table 14 intuitively reflect these trends. For example, tech-related engineering occupations are highly scored by OPENWEB, and these tend to have prestige scores over 60.

G Geographic locations

G.1 Geoparsing annotation & evaluation

For annotation of geographic locations, we verify, correct, or add to Mordecai3's predictions. Mordecai3 links mentioned locations to unique IDs in the GeoNames geographic database. We divided 200 randomly sampled ABOUT pages among authors to annotate, and follow annotation instructions shown in Figures 12 and 13. We use the context surrounding a mentioned location and pragmatic principles when making judgements, especially when exact locations are underspecified (Grice, 1975). To calculate interannotator agreement, 35 of 200 pages were doubly annotated. For assigning GeoName IDs, our annotators achieve high pairwise agreement on spans (Cohen's $\kappa = 0.809$). We also annotate whether subjects may *identify with* mentioned locations on their ABOUT page. Our agreement on this binary task is lower than that of GeoName IDs, likely due to the more subjective and interpretive nature of the task (Cohen's $\kappa = 0.652$).

Table 20 outlines the performance of the geoparser we apply onto ABOUT pages. Overall performance is hurt by imperfect recall of spans and accuracy on our data is lower than the country level accuracy of 94.2% reported in the Mordecai3 paper, which mostly trained and evaluated on news and Wikipedia data (Halterman, 2023). By aggregating results to the most frequent country at the page-level, we are able to better navigate errors that may occur at more granular levels. Still, we encourage future work to continue improving geoparsing performance, especially for a wide range of textual domains. Table 19 shows a more extensive overview of count statistics for countries, subregions, and regions present in AboutMe. Out of all pages, 79.5% identify with the most frequent country geoparsed from locations on the page. We also considered taking the country geoparsed from the first span of the page, but only 65.2% of pages were labeled to identify closely with this country.

G.2 Country metadata

For our analyses, we incorporate the following metadata for countries: continental region, subregion, gross domestic product (in USD), and anglophone status.

Continental regions and subregions. We use regions and subregions delineated by the United Nations's "Standard Country or Area Codes for Statistical Use."[13]. We add Taiwan to Eastern Asia and Kosovo to Southern Europe, as Mordecai3 produces these country codes. We exclude Antarctica from analysis, as there are less than three hundred webpages geoparsed to it. Since the UN subregions Polynesia, Melanesia, and Micronesia are infrequent in AboutMe, we group them into a single subregion of *Pacific Islands*.[14] This way, all included subregions contain at least 4k websites (Table 19).

Gross domestic product (GDP). Following Zhou et al. (2022), we gather GDP for each country from the World Bank.[15] We take the value from most recent listed year where GDP in USD is recorded, which is typically 2022 or 2021.

[12] Though *baby* showing up as a self-identified role may seem unusual, it occurs in contexts such as *I'm an 80s baby*.

[13] https://unstats.un.org/unsd/methodology/m49/

[14] https://www.britannica.com/place/Pacific-Islands

[15] https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

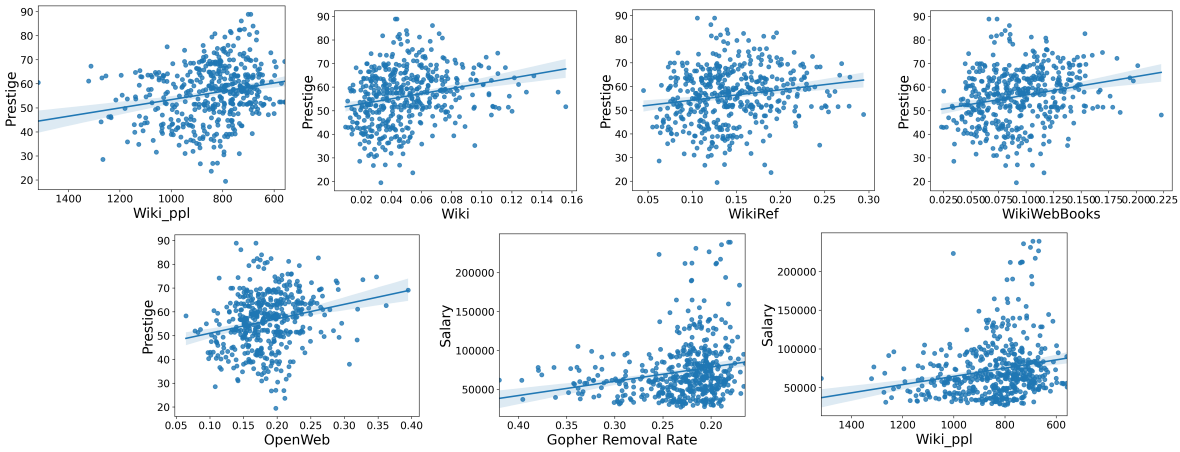

Figure 10: For some filters, we find statistically significant relationships between an occupation's prestige or salary (y-axes) and average filtering scores (x-axes), $p < 0.001$.

Anglophone status. The concept of an "English-speaking" country can be defined in a variety of ways. Official adoption of English does not necessarily entail high frequency of English use in a country, and vice versa (Plonski et al., 2013). For example, the United States has no official language, yet has a large majority of English speakers. Central to theories around the English-speaking world is that a few countries make up the "core anglosphere": the United States, Canada, the United Kingdom, Australia, and New Zealand (Vucetic, 2020). We bucket countries into four categories: "core" anglophone, English is an official and primary language, English is an official but not primary language, and all others. We use information about countries' official and primary language status aggregated on Wikipedia.[16]

G.3 Additional filtering results

We limit country-level filtering analyses only to countries that appear at least 500 times in our dataset, to ensure the patterns we find are over enough samples. We find weak Pearson correlations ($p < 0.05$, with Bonferroni correction) between a country's GDP and their average filtering scores for fastText, CLD3, and WIKI$_{ppl}$. However, these results are only due to a single outlier, China, which is often the most filtered country but also very high in GDP. After removing this outlier, all p values are insignificant, and thus we do not confidently conclude any broad relationship between wealth status and filtering.

In addition, we observe considerable overlap in filtering scores across the four levels of "English-speaking" countries (Appendix G.2). This finding, which is contrary to our hypothesis that filters may favor English-speaking locations, suggests that other factors may be at play aside from geography. For example, the topic *travel,tours* is the most common cluster (9.03%) of websites associated with Northern Africa, and travel websites may be written for outsider audiences and not reflect local communication patterns. Indeed, as our results in Appendix H show, topic usually has a higher and more significant influence on filtering scores than geography-related features.

H Regression

In §4.5, we run nine ordinary least squares regressions, one for each model-based filter, to investigate how different aspects of websites that we extract relate to filtering scores. To transform categorical variables into dummy binary variables, we use Africa as the base category for region, and *art, gallery* as the base category for topical interests. Since the directionality of how WIKI$_{ppl}$ should be interpreted is the opposite of other filters' scores, we negate its scores before performing its regression. Tables 21-29 show the results of these regressions in more detail. For clarity of interpretation, we include coefficients for only a subset of all topics with the most positive and negative effects in each regression. The topics with highest and lowest coefficients tend to reflect ones that are highly retained or removed by a filter.

[16] https://en.wikipedia.org/wiki/List_of_countries_and_territories_where_English_is_an_official_language

Tagged Role	Count
member	409814
artist	311808
director	298903
designer	232990
photographer	188463
founder	178863
teacher	176679
writer	174546
coach	168271
manager	151893
author	144552
owner	130686
president	130663
consultant	121052
editor	112568
student	92972
co	92363
engineer	88820
professor	87751
person	87704
instructor	87401
agent	85943
producer	85921
therapist	83870
realtor	80589
developer	79805
leader	79472
trainer	77860
professional	77430
mother	76335
speaker	76242
specialist	70193
mom	68985
graduate	67139
expert	66704
practitioner	65560
entrepreneur	60887
officer	59522
educator	58998
assistant	58078
musician	56813
ceo	54595
singer	54109
wife	53370
fellow	46894
girl	46476
lover	46279
native	45831
songwriter	45822
partner	44811

Table 16: The top 50 most frequently social role heads extracted by our RoBERTa token classifier.

Occupation family	Count	Examples of extracted roles
Arts, Design, Entertainment, Sports, & Media	1.1M	*artist*, *director*, *designer*, *writer*, *photographer*, *musician*, *player*
Production	620K	*designer*, *engineer*, *maker*, *builder*, *operator*, *mechanic*
Community & Social Service	452K	*therapist*, *educator*, *advisor*, *pastor*, *activist*, *social worker*
Computer & Mathematical	365K	*engineer*, *developer*, *scientist*, *strategist*, *programmer*
Educational Instruction & Library	308K	*teacher*, *professor*, *lecturer*, *curator*, *tutor*, *graduate student*
Healthcare Practitioners and Technical	300K	*therapist*, *nurse*, *doctor*, *nutritionist*, *surgeon*, *midwife*
Management	291K	*president*, *manager*, *dean*, *administrator*, *medical director*
Architecture and Engineering	288K	*architect*, *technician*, *electrical engineer*, *technologist*, *tester*
Business and Financial Operations	250K	*analyst*, *accountant*, *marketer*, *investor*, *management consultant*
Personal Care and Service	205K	*trainer*, *yoga teacher*, *stylist*, *makeup artist*, *caregiver*

Table 17: Ten most common O*NET occupation families in AboutMe, by website count, with example social roles. This is an extended version of Table 2.

Filter	Salary	Prestige
fastText	0.0554	-0.0178
CLD2	-0.0007	-0.0699
CLD3	0.1302	0.0339
langdetect	0.1353	0.0833
WIKI$_{ppl}$	0.2102***	0.2176***
WIKI	0.1051	0.2336***
WIKIREFS	0.1076	0.1713**
OPENWEB	0.1349	0.2280***
WIKIWEBBOOKS	0.1115	0.2238***
Gopher	0.2071***	0.1139

Table 18: Pearson correlation values between prestige or salary and filters' scores (higher = less filtered). For Gopher, we use the negated rate of page removal as the "score" since that filter does not output a single numerical score. We also negate scores for WIKI$_{ppl}$ so that its values can be interpreted similarly to other rows, since higher perplexities values get filtered more rather than less. Significance: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, with Bonferroni correction for 20 comparisons.

Country	Count
United States	3.0M
United Kingdom	803K
India	335K
Canada	306K
Australia	269K
China	139K
Germany	78K
New Zealand	78K
Italy	74K
South Africa	70K
Ireland	54K
France	52K
Netherlands	48K
Spain	47K
Japan	44K
United Arab Emirates	33K
Turkey	32K
Singapore	31K
Malaysia	31K
Nigeria	30K
Subregion	**Count**
Northern America	3.3M
Northern Europe	951K
Southern Asia	419K
Australia and New Zealand	347K
Western Europe	241K
Eastern Asia	237K
Southern Europe	204K
Sub-Saharan Africa	203K
South-eastern Asia	161K
Western Asia	155K
Latin America and the Caribbean	134K
Eastern Europe	118K
Northern Africa	21K
Pacific Islands	9.0K
Central Asia	4.6K
Region	**Count**
Americas	3.4M
Europe	1.5M
Asia	977K
Oceania	357K
Africa	224K

Table 19: The 20 most frequent countries in `AboutMe`, and ordered frequencies of all continental regions and subregions.

Task	Performance
Location span detection	P = 0.884, R = 0.768
Geoname IDs (all spans)	A = 0.627
Geoname IDs (recalled spans)	A = 0.795
Country (all spans)	A = 0.652
Country (recalled spans)	A = 0.826
Country (page-level)	A = 0.910

Table 20: Metrics showing how Mordecai3 performs on our dataset. Page-level country accuracy is determined based on whether the resulting country we link to pages is validly geoparsed from any location on the page. Key: P = precision, R = recall, A = accuracy.

Dependent variable: **WIKIWEBBOOKS**

Feature	Coefficient
Intercept	-1.170***
Topic: *news, media, content*	0.414***
Topic: *film, production, festival*	0.319***
Topic: *writing, books, book*	0.206***
Topic: *music, band, musical*	0.176***
Topic: *research, university, science*	0.152***
...	
Topic: *service, cleaning, repair*	-0.567***
Topic: *hair, beauty, skin*	-0.524***
Topic: *insurance, care, dental*	-0.504***
Topic: *home, homes, family*	-0.499***
Topic: *estate, real, property*	-0.476***
Region: Americas	0.090***
Region: Asia	0.002
Region: Europe	0.078***
Region: Oceania	0.074***
Individual	0.123***
Core anglophone	-0.147***
$\log_2$(# of characters)	0.142***
R^2	0.124
adj. R^2	0.124

Table 21: Regression results for the quality filter WIKIWEBBOOKS. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.

Dependent variable: **OPENWEB**

Feature	Coefficient
Intercept	-0.803***
Topic: *news, media, content*	0.772***
Topic: *people, world, work*	0.359***
Topic: *software, data, development*	0.315***
Topic: *writing, books, book*	0.315***
Topic: *like, love, time*	0.255***
...	
Topic: *service, cleaning, repair*	-0.387***
Topic: *estate, real, property*	-0.385***
Topic: *quality, equipment, production*	-0.353***
Topic: *home, homes, family*	-0.342***
Topic: *insurance, care, dental*	-0.302***
Region: Americas	0.104***
Region: Asia	-0.003
Region: Europe	0.089***
Region: Oceania	0.063***
Individual	0.088***
Core anglophone	-0.072***
$\log_2$(# of characters)	0.080***
R^2	0.077
adj. R^2	0.077

Table 22: Regression results for the quality filter OPENWEB. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.

Dependent variable: **WIKIREFS**

Feature	Coefficient
Intercept	-0.917***
Topic: *news, media, content*	0.490***
Topic: *club, members, association*	0.363***
Topic: *music, band, musical*	0.347***
Topic: *film, production, festival*	0.329***
Topic: *research, university, science*	0.253***
...	
Topic: *service, cleaning, repair*	-0.546***
Topic: *hair, beauty, skin*	-0.46***
Topic: *home, homes, family*	-0.414***
Topic: *furniture, jewelry, quality*	-0.411***
Topic: *products, quality, product*	-0.406***
Region: Americas	-0.006*
Region: Asia	-0.029***
Region: Europe	0.019***
Region: Oceania	-0.010***
Individual	-0.031***
Core anglophone	-0.049***
$\log_2$(# of characters)	0.114***
R^2	0.099
adj. R^2	0.099

Table 23: Regression results for the quality filter WIKIREFS. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.

Dependent variable: **WIKI**

Feature	Coefficient
Intercept	0.186***
Topic: *film, production, festival*	0.181***
Topic: *music, band, musical*	0.117***
Topic: *research, university, science*	0.112***
Topic: *club, members, association*	-0.007
Topic: *news, media, content*	-0.097***
...	
Topic: *blog, like, love*	-0.487***
Topic: *service, cleaning, repair*	-0.468***
Topic: *hair, beauty, skin*	-0.466***
Topic: *life, yoga, help*	-0.447***
Topic: *like, love, time*	-0.443***
Region: Americas	0.029***
Region: Asia	0.019***
Region: Europe	0.041***
Region: Oceania	0.039***
Individual	0.056***
Core anglophone	-0.165***
$\log_2$(# of characters)	0.016***
R^2	0.036
adj. R^2	0.036

Table 24: Regression results for the quality filter WIKI. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.

Dependent variable: **WIKI$_{ppl}$**

Feature	Coefficient
Intercept	-1.316***
Topic: *law, legal, firm*	0.121***
Topic: *god, church, christ*	0.118***
Topic: *insurance, care, dental*	0.075***
Topic: *research, university, science*	0.071***
Topic: *financial, clients, investment*	0.07***
...	
Topic: *online, store, shopping*	-0.38***
Topic: *fashion, women, brand*	-0.375***
Topic: *products, quality, product*	-0.341***
Topic: *quality, equipment, production*	-0.296***
Topic: *furniture, jewelry, quality*	-0.288***
Region: Americas	0.047***
Region: Asia	-0.102***
Region: Europe	0.063***
Region: Oceania	0.015***
Individual	0.041***
Core anglophone	0.002
$\log_2$(# of characters)	0.132***
R^2	0.078
adj. R^2	0.078

Table 25: Regression results for the quality filter WIKI$_{ppl}$. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.

Dependent variable: **fastText**

Feature	Coefficient
Intercept	-2.154***
Topic: *law, legal, firm*	0.310***
Topic: *insurance, care, dental*	0.292***
Topic: *children, child, school*	0.282***
Topic: *god, church, christ*	0.276***
Topic: *financial, clients, investment*	0.264***
...	
Topic: *online, store, shopping*	-0.37***
Topic: *quality, equipment, production*	-0.293***
Topic: *fashion, women, brand*	-0.262***
Topic: *products, quality, product*	-0.239***
Topic: *com, www, https*	-0.177***
Region: Americas	-0.078***
Region: Asia	-0.152***
Region: Europe	-0.004
Region: Oceania	-0.057***
Individual	0.073***
Core anglophone	0.112***
$\log_2$(# of characters)	0.206***
R^2	0.175
adj. R^2	0.175

Table 26: Regression results for the English filter fastText. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.

Dependent variable: **CLD2**

Feature	Coefficient
Intercept	-0.211***
Topic: *solutions, technology, business*	0.121***
Topic: *marketing, digital, media*	0.098***
Topic: *insurance, care, dental*	0.098***
Topic: *financial, clients, investment*	0.096***
Topic: *services, service, clients*	0.094***
...	
Topic: *quality, equipment, production*	-0.076***
Topic: *com, www, https*	-0.035***
Topic: *fashion, women, brand*	-0.001
Topic: *company, products, quality*	-0.001
Topic: *online, store, shopping*	0.005
Region: Americas	-0.047***
Region: Asia	-0.164***
Region: Europe	-0.050***
Region: Oceania	-0.059***
Individual	0.011***
Core anglophone	0.098***
$\log_2$(# of characters)	0.015***
R^2	0.009
adj. R^2	0.009

Table 27: Regression results for the English filter CLD2. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.

Dependent variable: **CLD3**

Feature	Coefficient
Intercept	-1.330***
Topic: *solutions, technology, business*	0.175***
Topic: *insurance, care, dental*	0.158***
Topic: *services, service, clients*	0.153***
Topic: *service, cleaning, repair*	0.148***
Topic: *financial, clients, investment*	0.146***
...	
Topic: *online, store, shopping*	-0.081***
Topic: *quality, equipment, production*	-0.075***
Topic: *fashion, women, brand*	-0.053***
Topic: *com, www, https*	-0.041***
Topic: *music, band, musical*	-0.012***
Region: Americas	-0.021***
Region: Asia	-0.174***
Region: Europe	-0.002
Region: Oceania	-0.033***
Individual	0.013***
Core anglophone	0.069***
$\log_2$(# of characters)	0.125***
R^2	0.061
adj. R^2	0.061

Table 28: Regression results for the English filter CLD3. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.

Dependent variable: **langdetect**

Feature	Coefficient
Intercept	-0.560***
Topic: *solutions, technology, business*	0.042***
Topic: *construction, project, projects*	0.034***
Topic: *services, service, clients*	0.031***
Topic: *children, child, school*	0.029***
Topic: *marketing, digital, media*	0.028***
...	
Topic: *online, store, shopping*	-0.069***
Topic: *fashion, women, brand*	-0.061***
Topic: *car, vehicle, auto*	-0.038***
Topic: *com, www, https*	-0.037***
Topic: *products, quality, product*	-0.031***
Region: Americas	-0.025***
Region: Asia	-0.071***
Region: Europe	-0.027***
Region: Oceania	-0.038***
Individual	0.014***
Core anglophone	0.053***
$\log_2$(# of characters)	0.055***
R^2	0.011
adj. R^2	0.011

Table 29: Regression results for the English filter langdetect. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.

Note: Some of the examples in our annotation task may be NSFW.

In this task, you will label occupations and roles that refer to the subject of biographies.

Step 0: Select files you've been assigned

Brat 🌭 is running at [link redacted]. Go to **roles_data** in the file system display screen ("Collections" button on the top left). After selecting a file assigned to you, **log in** using a username and password provided to you. Each file contains a sentence-long excerpt from a different *about* page.

Use the right/left arrows on your keyboard to quickly move between docs.

Step 1: Correct highlighted spans

Our definition of "roles" or "occupations" on *about* pages is any **singular noun referring to the subject of the bio**. If a span is highlighted and does not fall into this definition (e.g. it is not a person, or refers to someone else), delete it. The roles and occupations can be ones that the subject actively participated in the past, e.g. *Throughout my life I have been a **teacher**, a startup **founder**, and a seashell **collector**.*

Subject of the bio

- First person biographies: the subject is *I*, *me*, *my*, *mine*.
- Third person biographies: assume the bio's subject is the main person referenced in the excerpt sentence.
- These biographies have been automatically detected to be about individuals, but there may be some noise from that, and some bios can contain extraneous content. If it is unclear who in the sentence is the individual subject of the bio, then do not highlight any of the spans in that sentence.

Positive examples

- I am a **chef**, **author**, and **mom** living in Virginia.
- As an award-winning **geologist**, Sebastian has given talks around the world.
- **Knitter**, **blogger**, & **dreamer**.

In the last example above, the sentence's relation to the subject of the bio is implied rather than stated.

Negative examples

- My **wife** loves beekeeping as well.
- Janice works hard to accommodate every **client.**

Step 2: Add additional spans

Please also highlight the *heads* of any other noun phrases in the sentence that fall under our definition of a stated role/occupation on an *about* page.

Figure 11: Instructions for social role annotation.

In this task, you will be presented with text *about* a personal or organization. Text spans corresponding to named locations were highlighted and annotated by a model. You are going to check, correct, and/or add to these predictions.

Step 0: Select files you've been assigned

We'll be using brat , which is running at [link redacted]. Go to **geoparse_data** in the file system display screen ("Collections" button on the top left). After selecting one of your assigned files to annotate, **log in** (top right button) using a username and password provided to you.

Now, for each highlighted span, do Step 1 & Step 2. Some pages will have no spans highlighted, which means the model thinks there's no named locations. In that case, skip to Step 3.

Step 1: Verify geonames ID

Check the model's predicted geoname ID by double-clicking on a span and going to the geoname link in its "Notes" box.

- **If it is not correct**, replace it with a correct one (e.g. https://www.geonames.org/**2988507**/paris.html) by clicking on the "GeoNames" link. You can also use Wikipedia or Google to help disambiguate.
    - Use the text to help check or disambiguate:
        - "*I was born in London and later moved to Oxford*" → Oxford, UK since London is probably the UK one too
        - *"I am from Oxford, Ohio"* → Oxford, Ohio, U.S.A, because this is likelier than the person being from both Oxford, UK and Ohio separately
        - "*I live in Paris with my cat*" → Paris, France rather than Paris, Idaho, U.S.A., due to Grice's maxims of conversation.
        - **Rule of thumb**: usually the correct entry in geonames will have the same name as the entity in the text, e.g. searching "England" in GeoNames might rank "Great Britain" or "New England" higher but you should scroll down to "England".
        - It is fine if the model predicted one of multiple equally valid options, e.g. name of a city == name of its metropolitan region. If you are adding a new URL, choose the top ranked valid option.
    - If you can't find the correct geoname ID, leave the Notes field blank.
- **If "Notes" says *None* or is empty,** the model did not predict a geoname ID. Either leave it as "None" if geonames indeed doesn't include this location or add the correct URL. Following past work, we'll only evaluate on named places that have a geoname ID.
- **If you are confused,** write "confused" in the "Notes" box

Figure 12: The first half of instructions for geoparsing annotation. See Figure 13 for the second half.

Step 2: Annotate type of association

How often does the page's person/org *identify* with the mentioned location? Our model can't predict this, but we'll annotate this at a high level to get a sense of it for our paper. How we define this distinction:

Identifying with a location:

- is from e.g. was born in, originates from, grew up in, spent substantial time in
  - *Born in **Miami**, Camila was drawn to music at an early age.*
- currently lives, operates, or works in e.g. headquartered in, offers services in, considers home
  - *Priyanka is a **Delhi**-based performer.*

Not identifying with a location:

- visiting a location e.g. a travel blogger may list trips
- referring to someone else belonging to a location e.g. their husband is from somewhere
- a location they plan to go to in the future

In ambiguous cases, just leave the default category of ***associated***. For example, some organizations may list employees who each have different locations each identifies with, but may not indicate if the organization overall identifies with them.

Step 3: Add missing place name spans

Some locations may be incorrectly highlighted or not highlighted at all. Add or edit these highlights and then follow Step 1 & Step 2.

Aim for high recall of geoname IDs that pertain to **geopolitical entities**: cities, states, provinces, and countries, but also add other named place names that appear in GeoNames. GeoNames has an expansive sense of "named place names" it could contain, but its coverage of non-geographic regions (e.g. landmarks or famous buildings like Golden Gate Park or Arc de Triomphe) depends on many European/U.S.-centric data sources it's pulling from.

Figure 13: The second half of instructions for geoparsing annotation. See Figure 12 for the first half.