The Literary Canons of Large-Language Models: An Exploration of the Frequency of Novel and Author Generations Across Gender, Race and Ethnicity, and Nationality

Paulina Toro Isaza
IBM Research
Yorktown Heights, NY
ptoroisaza@ibm.com

Nalani S. Kopp
Ascend Consulting Global LLC
Brooklyn, NY
info@nalanikopp.com

Abstract

Large language models (LLMs) are an emerging site for computational literary and cultural analysis. While such research has focused on applying LLMs to the analysis of literary text passages, the probabilistic mechanism used by these models for text generation lends them to also understanding literary and cultural trends. Indeed, we can imagine LLMs as constructing their own "literary canons" by encoding particular authors and book titles with high probability distributions around relevant words and text. This paper explores the frequency with which certain literary titles and authors are generated by a selection of popular proprietary and open-source models and compares it to existing conceptions of literary canon. It investigates the diversity of author mentions across gender, ethnicity, nationality as well as LLMs' ability to accurately report such characteristics. We demonstrate that the literary canons of popular large-language models are generally aligned with the Western literary canon in that they slightly prioritize male authors and overwhelmingly prioritize White American and British authors.

1 Introduction

Large language models (LLMs) are an emerging site for computational literary and cultural analysis. Such research typically covers methods and evaluations for applying LLMs to creative writing (Gómez-Rodríguez and Williams, 2023) and literary analysis (Piper and Bagga, 2024) or for exploring the extent to which these models have been trained on partial or full literary texts (Chang et al., 2023). However, the probabilistic mechanism used by these models for text generation (Chang et al., 2024) lends them to use for also understanding literary and cultural trends. Indeed, we can imagine LLMs as constructing their own "literary canon" that form the basis of downstream tasks centered around literature such as recommendation, classification, and question-answering.

Traditionally, the debate about inclusion of texts in the canon has been held in undergraduate Literature departments when determining core curriculum. The debate, rooted in the heterogeneous definitions of "classics" [1] and "canon," [2] becomes more convoluted in literary criticism in the last century. The adjective "classics" evolved from signifying Greco-Roman antiquity "of the first class, or the highest rank or importance" to indicating a more general representation of art and literature over the past three centuries (Oxford English Dictionary, 2024b). It was not until 1929 where *Literary Criticism* was appended to the entry for "canon" relating the noun directly to a "body of literary works traditionally regarded as the most important, significant, and worthy of study; those works of esp. Western Literature considered to be established as being of the highest quality and most enduring value; the classics (now frequently in the canon)" (Oxford English Dictionary, 2024a). Richard Ohmann's definition furthers that the canon is a "shared understanding of what literature is worth preserving" (1983). Alternatively, Guillory (1987) discusses how maintaining the canon as a static form of representation is problematic, given that it inherently includes the elite, while further excluding social groups without power.

In our study, we consider if large language models "preserve" specific works of literature by encoding them with high probability distributions around relevant words and text. We suspect that because of repetition bias in model training and data set quality limitations, LLMs may proliferate the marginalization of specific marginalized demographics and the solidify the elite in literature. As people rely more on LLM-powered assistants or search engines to discover literary works, it becomes imperative to understand how these models generate recommendations.

[1] See Appendix A.1

[2] See footnote 1.

This paper explores the frequency with which certain literary titles and authors are generated by a selection of proprietary and open source models and compares them to existing conceptions of the literary canon. As mentioned above, the Western canon has long prioritized works written by male, white European authors with relatively recent strong push-back from Post-colonial and Feminist critiques (Morrissey, 2005; Gugelberger, 1991; Robinson, 1983). Following these critical perspectives, we analyze the diversity of author mentions across gender, ethnicity, and nationality. Additionally, we investigate the extent to which LLMs can produce accurate demographic information of authors.

This paper brings several novel contributions to the field of cultural analysis of large-language models:

1. An analysis of the most frequently mentioned titles and authors in English-language prompts about general fiction and literary canon across popular proprietary and open-source models. This includes cross-sectional analyses by author demographics of gender, race/ethnicity, and nationality.

2. An evaluation of model accuracy in producing the gender, race/ethnicity, and nationality of authors.

3. An open-source dataset of author demographics including gender, race/ethnicity, and nationality.

4. A data-driven analysis that confirms the LLMs' output further emphasizes a White, male, Western literary canon.

2 Related Work

2.1 Literary Canon

Literary critics acknowledge the troublesome nature of the formation of the canon. Guillory (1987) discussed how social determinants impact measuring the qualitative "value" of texts included in the traditional Western literary canon. Value can be measured as "representative of a given constituency" in an anthropological sense or as an "aesthetic artifact" typically confined to an elite class. He contested, along with other scholars, William Bennett's valuation of the canon as homogeneous.

In "To Reclaim a Legacy", Bennett (1984) asserted the importance of a shared cultural heritage and criticized non-Western and contemporary works' inclusion in the canon. Bennett states his purpose as creating a representative canon that reflects Western culture - a culture which he clearly views as being exclusively male and White, except for the token nineteenth-century representatives of Austen, Eliot and the twentieth-century representative of MLK, Jr (Appendix A.2 Table 8). In fact, Bennett's canon is 84% male, 96% White, 8% Latino, 92% Western European, and 8% South American. Similarly, Bloom (1994) evaluated twenty-six similar canonical works on the basis of aestheticism (Appendix A.2 Table 9). He argued that "resenters," such as Feminist and Post-colonical critics, were displacing their guilt by adapting the canon to suit their sociopolitical agendas and that we should also abandon readers who are "amenable to a politicized curriculum" (Bloom, 1994). This archaic perspective reinforces that education is limited to the elite both as lecturer and as student. Not surprisingly, Bloom's canon's distribution is 95% male, 5% female, 97% White, 3% Black, 66% Western European and 34% American.

The formation of a global literary canon is just as contested as that of a Western canon. The phrase "world literature" is credited to Goethe who criticized the narrowness of focusing on only one's national literature (i.e., canon) (Damrosch, 2003). Damrosch gives a more formal definition of world literature as dynamically "[circulating] out into a broader world beyond its linguistic and cultural point of origin". This particular definition allows Western works to sit as a subset of the global canon. Meanwhile, when determining core undergraduate curricula, American institutions frequently separate their introductory literature survey (aligned to the Western Canon) and their World/Comparative literature survey (global literary canon). In either perspective, the global literary canon is not meant to be merely a copy of the Western canon with only a few token non-American and non-European additions.

There have been attempts to broaden the Western literary canon beyond these perspectives, as illustrated by the addition of authors such as Richard Wright, Zorah Neale Hurston, Maxine Hong Kingston, and Junot Diaz in the IB high school literature curriculum (International Baccalaureate Organization, 2025). However, we should not make the mistake of thinking that the exclusive perspectives of Bennett and Bloom are a relic of the past. In 2022, the American Library Association reported a 38% annual increase of attempts to ban particular books in U.S. schools, the majority of which "written by or about members of the LGBTQ+ community and people of color" (2023). Such narrow catalogs and censorship impose a limited, elite perception of history and literary aesthetics on a diverse student population instead of reflecting the reality of a globalized world (Guillory, 1987).

Our study aims to investigate if LLMs fulfill a similar role in imposing such a view on a global, diverse set of users. Given the skewed demographics of AI professionals in which only 18% are female and 5.6% Black or Latino (when considering American Ph.D graduates), one can imagine a new elite class that controls the creation of large-language models (Zhang et al., 2021). We hypothesize that LLMs' outputs regarding literary canon will disproportionally represent this elite class who train them as "bias in AI can arise from different stages of the machine learning pipeline, including data collection, algorithm design, and user interactions" (Ferrara, 2023). Upstream inherited bias then flows downstream with real-world impacts, such as text-to-image models like StableDiffusion, DALL-E, and Midjourney mirroring the under-representation of female CEO's and associating people of color with criminals or terrorists (Mittelstadt et al., 2016). For LLMs, AI professionals select which authors get represented in models through their choice of training corpora. These choices can be explicit such as intentional decisions about which books are included or implicit such as including corpora collected by others without concern for the bias that might exist in such data collections. We believe that the current lack of diversity in these professionals will inevitably contribute to downstream bias in applications conducting computational literary analysis.

2.2 Computational Literary Analysis

Computational literary analysis, situated within the digital humanities, has long made use of computational methods to analyze literary narrative text. Examples include BookNLP's named entity extraction and co-reference resolution for character, supersense, and event analysis (Bamman et al., 2014), character network extraction and analysis (Labatut and Bost, 2019), and linear regression with TF-IDF and doc2vec embeddings for detecting the degree of narrativity in a given passage (Steg et al., 2022). It is also worth noting the use of computational techniques in literary analysis is not without its critics (Da, 2019).

Recently, generative large language models have been added to this repertoire. For example, Piper and Bagga (2024) used various open-source and proprietary large-language models to capture more than a dozen narrative features from literary passages across point of view, time, and setting. In another study, Yu et al. (2024) created a dataset for evaluating large-language models on questions about Chinese literary text, finding that even large models like ChatGPT struggle with answering questions regarding literary aspects such as character, style, and plot.

Another avenue of research has focused on investigating which exact texts were used in training which is known to frequently leverage literary texts (Chang et al., 2023). Such third-party investigations are imperative because the developers of LLMs do not typically publicize their training data; at most, they might only mention some of their high-level datasets. Chang et al. (2023) show that GPT-4 is more likely to intimately know works in the public domain in the U.S., genre science-fiction, and fantasy novels. To a lesser extent, it knows a bit of about horror, thrillers, and general bestsellers. It is least likely to have been trained on Anglophone fiction written outside of the U.S. and U.K. as well as works by Black authors.

We expand on this work by changing the scope from full texts used to train models to investigating the models' general awareness of different titles and authors. This does not require a model to be trained on the full text but rather any text that mentions the author and book title such as Wikipedia, reviews, discussion forums, and literary criticism.

3 Methodology

3.1 Models

The study evaluated both propriety and open-source models listed in Table 1. Most models were of relatively large size, with only one small model of eight billion parameters. While the number of parameters of GPT 4o and Gemini 1.5 Pro are not published, both models are much larger than the Llama models tested here.

Model	License
GPT 4o	Proprietary
Gemini 1.5 Pro	Proprietary
Llama 3.3 70B-Instruct	Open-Source (Custom)
Llama 3.1 8B-Instruct	Open-Source (Custom)

Table 1: Models evaluated on the book title generation task.

3.2 Book Title Generation

This first experiment generated title and author pairs with a variety of prompts for use in the subsequent steps in the methodology (see Table 2). For more information about the model parameters and post-processing, see Appendix B.1.

Models were prompted to generate varying amounts of book titles both with and without providing a more specific description of the type of literary canon requested. The prompts tested included the following descriptive: no description, “fiction”, “classic”, “literary canon”, ”Western literary canon”, and “global literary canon”. Using the descriptive“literary canon,” we reviewed the output’s correlation with previous definitions of “canon” and “classics” and the extent to which the LLM considered Western literary canon as default. By specifying our prompts to recommend works from the “western literary canon” and “global literary canon,” we tested if the LLM produced a more diverse set of authors and titles. The more general descriptive of “fiction” and the blank descriptive were used as baselines.

For each description, the models were separately prompted to generate 5, 10, 20, 50, and 100 samples. Multiple prompting styles shown in Table 2 were tested to ensure that results were not unique to a specific prompt. Additionally, a few variations of one of the prompt styles (#1.1) were used to force the model to separately generate older and more contemporary titles.

3.3 Author Demographic Generation

In this second experiment, the models were prompted to generate select demographic information (gender, race/ethnicity, and nationality) for the purpose of evaluating the models’ ability to correctly output such data. Parameters and post-processing methods are reported in Appendix B.3.

The prompt styles for generating the author demographic information were designed to prompt the models to mimic a lay person’s casual interactions with such a model (Table 3). For this reason, no definitions or limitations of the particular demographic feature were provided. Additionally, no instructions for output format were given, in order to minimize results with errors producing the wrong format with the right information.

3.4 Human Annotations

The ground truth annotations formed the basis of the demographic and publishing information of the LitAuthorDemoDB dataset presented in Section 4. The two researchers manually created labels for each author’s gender, race/ethnicity, and nationality. Race and ethnicity categories were based on race categories from the U.S. Census along with the additional suggested MENA (Middle Eastern or North African) category and the Hispanic/Latino ethnic question. The individual labels were chosen based on the author’s Wikipedia page, their official website, and interviews.

Extra care was taken in cases where an author carried multiple citizenship or identified with multiple nationalities. However, this information was not always readily available and the authors (as persons with dual-citizenship themselves) recognize that nationality can be more nuanced than captured in tabular data. For this reason, each author recorded included a “Notes” column which is available in the open-source LitAuthorDemoDB.

Additionally, as White is often considered the default, many authors who might identify as White did not have this identity explicitly stated in biographies or interviews the way authors of other races typically do. The annotators used the White label for race/ethnicity if the author did not claim any other identity and appeared white passing. This is a problematic and imperfect annotation rule, but it was determined to result in more accurate information than the alternative of leaving the majority of White authors without a label.

Inter-annotator agreement was evaluated using Cohen’s Kappa coefficient by comparing the three annotation categories across 100 randomly sampled authors. The coefficient for gender was 1 with all labels matching. The coefficient for race and ethnicity was lower at 0.90 with variances arising mostly from authors with multiple racial and ethnic identities. The agreement for nationality was the lowest with a coefficient of 0.76. Disagreements typically involved authors who were first and second generation immigrants with labels sometimes but not always including the author’s birth or their

ID	Prompt
1.1	Name [n] [descriptive] books
1.1.1	Name [n] [descriptive] books published before 2000
1.1.2	Name [n] [descriptive] books published after 2000 and before 2015
1.1.3	Name [n] [descriptive] books published after 2015
1.2	Recommend [n] [descriptive] books
1.3	Can you recommend [n] [descriptive] books?
1.4	What [descriptive] books should I read?
1.5	What are the [n] best [descriptive] books?

Table 2: Prompts used for the book title generation task. The values for **descriptive** were: “fiction”, “classic”, “literary canon”, “Western literary canon”, “global literary canon”, and blank. Values of **n** were 5, 10, 20, 50, 100.

ID	Prompt
2.1	What is author [name]’s gender?
2.2	What is author [name]’s race/ethnicity?
2.3	What is author [name]’s nationality?

Table 3: Prompts used for the author demographic generation task.

parents’ birth country. Other disagreements occurred for authors from the UK who were labeled British by one annotator and English by another. A non-systemic peer review resolved some but certainly not all of these discrepancies.

4 AuthorDemoDB

We present LitAuthorDemoDB, an open-source dataset of classic and contemporary authors with corresponding demographic information including gender, race/ethnicity, and nationality. While author datasets such as Gale’s Books and Authors database and ISBNdb exist, they are not easily or freely accessible. Indeed, there is no direct download of datasets or API access to easily match an author to demographic information. LitAuthorDemoDB is meant to provide readers and researchers an accessible, open-source, and community-updated and reviewed repository for author demographics. It is available for download at https://github.com/IBM/LitAuthorDemoDB. We plan to continually update with new authors and fields, particularly to increase the diversity of the dataset.

The first version of the dataset contains a total of 1,345 authors and 2,238 corresponding titles. In Appendix C, Table 12 shows the author and book table schema. The current dataset is majority male (58%) with 41% female authors and 11 non-binary authors. It also contains a majority of White authors (78%). Meanwhile, 8% of authors are Asian, 8% are Black, and 3% are Latino. At least one, but less than 1% of authors are of the following racial and ethnic categories: Native American, Pacific Islander, and Aboriginal Australian. While the authors represent seventy-nine nationalities, about half of the authors are American and 20% are British. All other nationalities account for less than 5% of the dataset.

The next version of the dataset will draw from a variety of genres as well as other sources such as WikiData with a focus on increasing gender, racial, and national diversity. Users will also be able to suggest corrections and new authors.

5 Experimental Results

5.1 Generated Titles and Authors

In total, the book title generation prompts described in Section 3.2 produced at total of 30,302 author and title pairs across the four models and various prompting styles. They generated 1,347 unique authors across 2,238 unique titles. When only considering the prompt styles invoking categories of literary canon, the dataset included 1,021 unique authors across 1,640 unique titles. Tables 13 and 14 in Appendix D show the distribution of unique authors and titles according to model.

The frequencies of titles and authors were highly skewed. The majority of titles were mentioned with a median of 2 but average of 13.5. This trend also held for author mentions with a median of 4 but average of 22.5. 69% of authors had only one book title while 6% had at least five titles associated. The author with the most number of works was Shakespeare.

5.1.1 Top Generated Titles and Authors

The ten most common generated author and title pairs are shown in Table 4. While there were slight variations in the top pairs by model, they generally overlapped in which titles were most mentioned. Interestingly, the single top generated title was the same for all four models tested: *Pride and Prejudice* by Jane Austen.

Title	Author	N
Pride and Prejudice	Jane Austen	652
The Great Gatsby	F. Scott Fitzgerarld	489
To Kill a Mockingbird	Harper Lee	443
1984	George Orwell	418
Don Quixote	Miguel de Cervantes	391
Jane Eyre	Charlotte Brontë	371
The Odyssey	Homer	357
Wuthering Heights	Emily Brontë	353
One Hundred Years of Solitude	Gabriel García Márquez	336
The Catcher in the Rye	J.D. Salinger	322

Table 4: Top 10 title and author pairs.

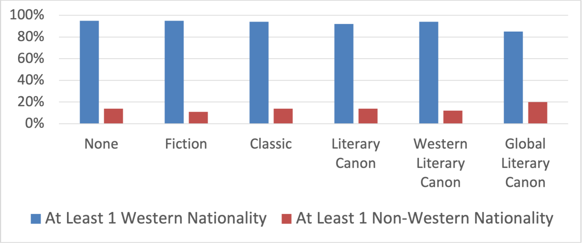

Figure 1: Distribution of authors with at least one Western nationality vs. authors with at least one non-Western nationality across prompt descriptions.

Prompting the model with different descriptors such as "fiction", "classic", or "global literary canon" only resulted in small variation between the titles generated. When offering no specifics about the type of book in the prompt, the models generated the largest number of distinct titles and the most divergent set of top ten titles. Along with the generic "fiction" descriptor, it was the only descriptor to generate popular literature in the top ten such as *The Lord of the Rings* and *The Girl with the Dragon Tattoo*. However, half of the top ten generated titles for no descriptor and the "fiction" descriptor were titles very firmly in the Western literary canon.

Figure 1 demonstrates that the overwhelmingly majority of authors generated across prompts had at least one Western nationality. It is only when considering "global" literary canon that we see an increase to 15% of generated authors coming from outside of the Western world. Alternatively, we can consider authors with single or dual nationalities of which at least one is outside of the U.S., Canada, Europe, and Australia. We see that such authors account for 14% of those generated by the "literary canon" prompt and 12% by the "Western literary canon" prompt. The proportion only increases to 20% for the "global literary canon" prompt. This behavior was consistent across all four models.

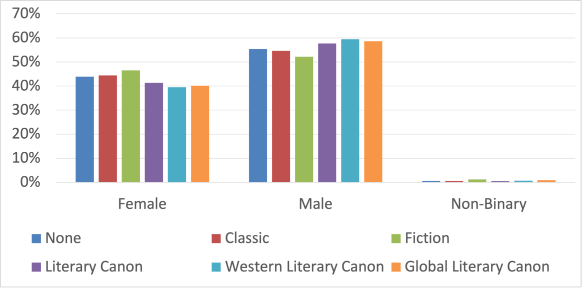

Figure 2: Distribution of author gender across prompt descriptions.

5.2 Author Demographic Distributions

We analyzed the distribution of authors across gender, race/ethnicity, and nationality. We discounted titles that were written by multiple authors (about 50 records) leaving 1,298 total authors.

When considering gender, no prompt description resulted in female authors accounting for half of the total output (Figure 2). The closest description was "fiction" of which 45% of the authors were women. This description also had the most non-binary authors at 5. The most male-skewed description was "Western literary canon" at 58% male although "global literary canon" was not far behind at 57%.

The gender distribution depended on the model used: the proportion of male authors ranged from 55% to 63% while the proportion of female authors ranged from 33% to 41% (Llama 3.1 8B and Llama 3.3 70B respectively). Gender also affected how often an author was mentioned: on average male authors were mentioned 1.6 times as often as female authors and 3.8 times as often as non-binary authors.

The distribution according to race and ethnicity was fairly stable no matter the description used to prompt the models as shown in Figure 3. White authors were the most represented across all description types. Asian, Latino, and Middle Eastern or North African saw a small increase in prompts for "global literary canon" compared to other descriptions but never broke past 12% of the authors generated.

Of the four models tested, only the smaller model, Llama 3.1 8B Instruct, varied substantially in the distribution of authors by race and ethnicity. In particular, it generated less Black (4%) and Asian (6%) authors and more White (82%) authors than the larger models. As with gender, an author's race and ethnicity influenced the rate at which an

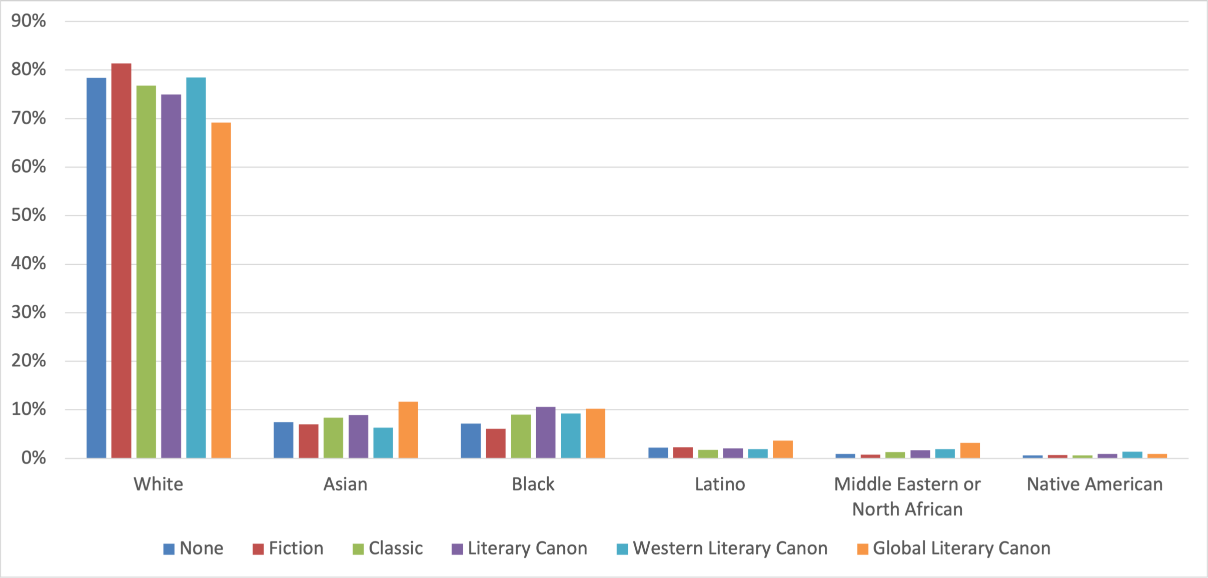

Figure 3: Distribution of author race and ethnicity across prompt descriptions. Pacific Islander and Aboriginal Australian omitted due to small sample size.

author's works were generated. On average, White authors' works were cited 1.5 times as often as those of Black authors, 1.8 times more than Middle Eastern or North African authors, 1.9 times more than Asian authors, and 2.5 times as Native American authors. Although substantially less Latino authors were cited in total, Latino authors' works were cited slightly more often than White authors.

All descriptions and models overwhelmingly favored authors from The United States (52%) and the United Kingdom (20%). See Table D in Appendix D for the full distribution of the 79 nationalities represented in the data. All other nationalities accounted for less than 5% of all authors, no matter the prompt description used. The models tended to produce similar distributions across nationalities. Only when specifying "global literary canon" did some nationalities outside of The United States, Canada, and Europe start to see increases, but these "Western" nations still made up the majority of the top. Only Japan, India, China, Nigeria, and Iran were able to account for more than 1% of authors even with this specification while the U.S. and the U.K. still accounted for more than 60% of authors.

However, unlike with gender and race and ethnicity, authors of the majority nationalities were not more likely to be mentioned. American authors on average had 6.85 mentions, placing it at 22nd place. While some nationalities that only accounted for a small percentage of the authors, those few authors' works were very popular with the models. For example, Colombian authors (representing 0.3% of all authors) had their works cited on average 35.6 times. English authors in particular both accounted for a large proportion of all authors (11.2%) and those authors whose work was regularly mentioned (average 15.6 times).

5.3 Evaluation of Model Generation of Author Demographics

We prompted the four models to generate the author's gender, race/ethnicity, and nationality. Overall, the models were generally able to accurately generate this information.

Models were most successful in generating the correct gender of an author (Table 5). GPT-4o and Llama 3.3 70B were the most accurate although Gemini 1.5 Pro was not far behind. The smaller model, Llama 3.18B, struggled with a number of female authors and output that it had no information about them.

Model	Female	Male	Non-Binary	Total
Llama 3.1 8B	0.85	0.95	0.91	0.91
Llama 3.3 70B	0.98	0.98	1.00	0.98
Gemini 1.5 Pro	0.97	0.97	0.73	0.97
GPT 4o	0.99	0.99	1.00	0.99

Table 5: Accuracy of author gender generation per model.

In Table 6 we report the recall of predictions of the positive class for each binary race and ethnicity flag. We choose to focus on recall because of some-what common false positives in the post-processing due to outputs including information about White authors writing about characters of other races and ethnicities or White authors who were born in former colonies. The three larger models performed similarly across race and ethnic categories, with slightly lower performances for Latino, MENA, and Native American authors. Interestingly, when the models failed to predict that the author was White, it was because they made no mention of the author's race or ethnicity. In many cases, they only referred to American or European nationality and did not differentiate between European nationalities and ethnicity. As with gender, when generating an author's race and ethnicity, Llama 3.1 8B struggled the most.

We report the recall of author nationality generations for similar reasons to race and ethnicity, particularly because of false positives of White authors born in former colonies. The recall for each model was relatively high with the smaller Llama 3.1 8B once again performing the lowest (Table 7). However, it's important to note that 89% of nationalities had less than twenty examples, with a little over half only having one or two examples. While the recall for these nationalities was still high, it is difficult to make generalizations of the models' performance on these nationalities based on such small samples.

6 Discussion

When analyzing the presence of bias or skewness of distributions, the question of what constitutes an unbiased distribution is not trivial. In the context of equitable literary representation of demographic groups in large language model generations, we can consider various distinct conceptions of a fair distribution. The first compares the distributions generated by LLMs to current existing distributions of established lists of literary canon. We can also use the distribution of the publishing industry as a comparison. Alternatively, we can compare the LLM distributions to actual demographic trends. The most strict definition compares against a uniform distribution of all possible demographic categories.

All four models exhibited a similar "understanding" of the concept of the literary canon. The large percentage of Western authors generated cross the phrases "literary canon", "Western literary canon", "global literary canon", and "classic" (Figure 1) suggests that these models default the literary canon to the Western literary canon. Indeed, they continue to prioritize Western works even when asked to consider the subject at a global scope. To illustrate this result, specifying the "global literary canon" only resulted in two of the top ten spots being held by authors that were not European or American: *One Hundred Years of Solitude* by Gabriel García Márquez (Colombia) and *The Epic of Gilgamesh* (Ancient Mesopotamia). In addition, 40% of all authors generated by the more generic blank and "fiction" prompts were also generated by the literary canon prompts. These findings suggest that model training data has been skewed heavily towards the Western canon. This bias can have broad implications for downstream tasks regarding literature and creative writing. Users will have to be explicit when prompting models if they want a broader range of output than the LLM's Western canon.

In regards to gender, the evaluated LLMs were substantially more diverse than the limited lists offered by Bennett and Bloom which were only 4-16% female. Meanwhile, female authors represented 41% of those generated by LLMs. The literary canon of LLMs is substantially more gender diverse than the more restrictive canons as well as earlier publishing trends until 1900 where women made up of about 10 % of published authors (Rosalsky, 2023). However, it still falls short of reflecting the the global gender distribution in which men (50.4%) slightly outnumber women (49.6%) (Carey and Hackett, 2022).

The racial, ethnic, and national demographics of generated authors across all prompt descriptors (including the generic "fiction" and blank descriptor) align to less inclusive catalogs of Western canon, created by critics such as Bennett and Bloom (Bennett, 1984; Bloom, 1996). When using ISBN registrations as a proxy for global publishing trends, American authors in the dataset are represented at similar rates of the global publishing industry share (both 52%) and British authors are represented at a vastly disproportional rate (20% vs. 3%) (World Intellectual Property Organization, 2022). Global publishing data concerning author race and ethnicity is not typically aggregated, in part because not all countries publish such data at the national level. Within the American publishing industry, it is estimated that 95% of authors published between 1950 and 2018 were White with the number increasing

Model	Asian	Black	Latino	MENA	Native American	Pacific Islander	White
Gemini 1.5 Pro	0.97	0.99	0.91	0.95	0.90	1.00	0.58
GPT 4o	0.96	0.98	0.97	0.91	0.90	1.00	0.34
Llama 3.1 8B	0.91	0.92	0.86	0.86	0.80	1.00	0.22
Llama 3.3 70B	0.99	0.97	0.97	0.95	0.90	1.00	0.62

Table 6: Recall of author race/ethnicity generation per model across binary race and ethnicity categories. MENA = Middle Eastern or North African. Authors could have multiple positive race/ethnicity flags.

Model	Recall
Gemini 1.5 Pro	0.96
GPT 4o	0.95
Llama 3.1 8B	0.87
Llama 3.3 70B	0.98

Table 7: Recall of nationality generation per model. Authors could have multiple positive nationality flags.

to 89% when only examining those published in 2018 (So and Wezerek, 2020). In comparison, the generated American authors were 77% White, suggesting that these LLMs are not always replicating disparate publishing trends.

In regards to population demographics, White male authors from the U.S. and U.K. are overrepresented in relation to regional and global demographics. For example, White American authors account for 77% of American authors versus 58% of the American population (Jensen et al., 2021). When considering authors of all nationalities, 74% identified as only White. While global demographic datasets compiled with such racial and ethnic categories are harder to come by, it is fairly clear that this 74% figure grossly over-represents the number of people who identify as White throughout the world.

Ultimately, our experiment results demonstrate that current popular large-language models generate output about literary titles and authors that is biased in comparison to population demographic baselines. However, these models sometimes reflect while other times opposing the biased trends of the global publishing industry or formalized lists of literary canon. We suspect that the demonstrated biases occur because of (English) pre-training text that overwhelmingly discusses a small range of authors. This is evidenced by the much higher average than median of mentions per title and author. The behavior around the "global" prompt also suggests that models are not learning to disentangle the hegemony of Western culture from the concept of a literary canon. The extent to which these behaviors are due to more explicit instruction-tuning or fine-tuning on biased labeled data is hard to determine. Even so, such tuning can be the source of bias mitigation for tasks around generating literary titles and authors.

7 Conclusion

Our evidence suggests that the literary canons of popular large-language models are generally aligned with common conceptions of the the Western literary canon in that they slightly prioritize male authors and overwhelmingly prioritize White American and English authors particularly in comparison to global population demographics. This behavior occurs even when explicitly prompting models for a broader 'global" canon. We advocate for a globalized representation of canonical standards within LLMs, using our dataset as a vehicle to align output to better reflect international demographics. We are concerned that ancient, historical, and contemporary texts from entire continents such as Africa and Asia and aboriginal and native cultures from the Americas account for less than nine percent of nationalities represented in the "LLM canon". Additionally, while LLMs appear to accurately reproduce demographic information, further study should be considered with concerns over personal identity and biographical fact. Other potential areas for further study include: prompting models in different languages; running experiments with different sampling parameters; investigating the diversity of popular and genre literature; including other demographic information such as LGBTQIA+ status; and evaluating the model's ability to complete more complex tasks such as question-answering of titles written by a diverse set of authors. We urge our readers to contribute to our LitAuthorDemoDB as our hope is to leverage it to re-train LLMs with a more diverse, representative canon, impacting future analysis, scholarship, and readership.

Limitations

There are several limitations to the current study including prompt language and design, model parameters, postprocessing methods, and annotation methods.

This preliminary paper limits its scope to English-language prompts which potentially inherently privileges English-speaking (correlating with Western) perspectives. Additionally, only testing models that were developed by US-based companies could enhance this bias. A natural next step would be to include prompts in other languages as well as test models developed in other regions.

The researchers did not carry out prompt engineering or use model-specific system prompts in order to evaluate the model generation in the most generic of contexts. Using recommended model-specific system prompts for chat assistants could have changed the output.

There was no systemic check of the postprocessing used to compare the demographic model predictions to the ground truth labels. More robust postprocessing for evaluating the generated demographic information would allow reporting accurate precision and F1 instead of only recall. The ground truth labels themselves were created by the two researchers with only a minimal number checked for inter-annotator agreement.

Ethics Statement

Many of the models employed in this study were most likely trained on copyright data. While this study is not meant to show end users how to replicate copyright data, the authors acknowledge that simply using the models might constitute harm to copyright holders. Additionally, the authors did not solicit third-party annotation but rather performed annotation themselves. However, as with copyright data, many of the models used were likely also trained using data created by underpaid and exploited human annotators, particularly in the global south.

There is unfortunately no standard way of assessing the environment cost of running model inference. The authors acknowledge that running such experiments with hundreds of prompts across multiple large models most likely contributed to substantial environmental cost including both direct costs and indirect costs such as increased demand for additional environment-damaging data centers.

Acknowledgments

We gratefully acknowledge the support of a few people in the making of this article: Yu Deng, Daby Sow, and Dr. Bobby Birhiray, M.D., M.T. Your insights were valuable in our discussion and recommendations for future study. Nalani would also like to express gratitude for her first daughter, Kesiena Solene Birhiray, for reminding her to follow her purpose.

References

American Library Association. 2023. American library association reports record number of demands to censor library books and materials in 2022. Accessed on February 23, 2025.

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379, Baltimore, Maryland. Association for Computational Linguistics.

W. J. Bennett. 1984. To reclaim a legacy: A report on the humanities in higher education.

Harold. Bloom. 1994. *The Western Canon: The Books and School of the Ages*. Harcourt Brace.

Harold Bloom. 1996. The western canon: The books and school of the ages. *History of the Human Sciences*, 9:99–99.

Isabel Webb Carey and Conrad Hackett. 2022. Global population skews male, but un projects parity between sexes by 2050. *Pew Research Center*. Accessed on March 22, 2025.

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.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Nan Z. Da. 2019. The computational case against computational literary studies. *Critical Inquiry*, 45(3):601–639.

David Damrosch. 2003. *What Is World Literature?* Princeton University Press.

Emilio Ferrara. 2023. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3.

Gale. Gale books and authors. Accessed on February 20, 2025.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: a comprehensive evaluation of LLMs on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14504–14528, Singapore. Association for Computational Linguistics.

Georg M. Gugelberger. 1991. Decolonizing the canon: Considerations of third world literature. *New Literary History*, 22(3):505–524.

John Guillory. 1987. Canonical and non-canonical: A critique of the current debate. *ELH*, 54(3):483–527.

International Baccalaureate Organization. 2025. Prescribed reading list. Accessed on February 24, 2025.

ISBNdb. Isbn database. Accessed on February 20, 2025.

Eric Jensen, Nicholas Jones, Megan Rabe, Beverly Pratt, Lauren Median, Kimberly Orozco, and Lindsay Spell. 2021. 2020 u.s. population more cacially and ethnically diverse than measured in 2010.

Vincent Labatut and Xavier Bost. 2019. Extraction and analysis of fictional character networks: A survey. *ACM Comput. Surv.*, 52(5).

Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):2053951716679679.

Lee Morrissey. 2005. *Debating the canon: A reader from Addison to Nafisi*. Springer.

Richard Ohmann. 1983. The shaping of a canon: U.s. fiction, 1960-1975. *Critical Inquiry*, 10(1):199–223.

Oxford English Dictionary. 2024a. canon, n. In *Oxford English Dictionary*. Oxford University Press. Accessed on February 20, 2025.

Oxford English Dictionary. 2024b. classics, n. In *Oxford English Dictionary*. Oxford University Press. Accessed on February 20, 2025.

Andrew Piper and Sunyam Bagga. 2024. Using large language models for understanding narrative discourse. In *Proceedings of the The 6th Workshop on Narrative Understanding*, pages 37–46, Miami, Florida, USA. Association for Computational Linguistics.

Lillian S. Robinson. 1983. Treason our text: Feminist challenges to the literary canon. *Tulsa Studies in Women's Literature*, 2(1):83–98.

Greg Rosalsky. 2023. Women now dominate the book business. why there and not other creative industries? *NPR*. Accessed on March 22, 2025.

Richard Jean So and Gus Wezerek. 2020. Just how white is the book industry? *The New York Times*. Accessed on March 22, 2025.

Max Steg, Karlo Slot, and Federico Pianzola. 2022. Computational detection of narrativity: A comparison using textual features and reader response. In *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 105–114, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

World Intellectual Property Organization. 2022. The global publishing industry in 2022.

Linhao Yu, Qun Liu, and Deyi Xiong. 2024. LFED: A literary fiction evaluation dataset for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10466–10475, Torino, Italia. ELRA and ICCL.

J. Zhang, I. Benaich, and Y. Shoham. 2021. *Artificial Intelligence Index Report 2021*, chapter Diversity in AI. Standford University.

A Literary Canon

A.1 Definitions and Etymology

- classics: The OED offers 2 entries for “classics,” with 15 definitions and 5 etymologies. The adjective form definition “of acknowledged excellence or importance” has inconsistencies in detail dating from 1597 to 2010. The main variance is whether “classical” requires a link to Greco-Roman antiquity or if it solely means “of the first class, of the highest rank or importance; constituting an acknowledged standard of model; of enduring interest and value” (see Adjective definitions I.1 versus I.2). Indeed when you turn to the etymology of Latin classicus the word originates from “class” (n.) indicating social standing relating to “groups, ranks, or categories” (entry I). This relates to the Middle French, French classique meaning of the highest rank with a reference in 1548 to medieval authors held in high esteem and 1680 to the best Latin authors.

- canon: The OED provides 7 entries for “canon,” with 192 definitions and 86 etymologies. The first entry originating in Old English indicates a connection to decrees of the Church, the second entry from 1588 links to “a general rule, fundamental principle... governing the systematic or scientific treatment of a subject; e.g. canons of descent or inheritance; ... canons of criticism, taste, art” (2.a.b). This definition relates to the changing meaning of “classics” to become more representative of people’s class or a body of work. Additionally, the etymology shows Latin canon meant rule (Etymology of “canon”). For our purposes, we selected the entry related to literary criticism.

A.2 Lists of Literary Canon

Author/Work
Homer
Sophocles
Thucydides
Plato
Aristotle
Vergil
Dante
Chaucer
Machiavelli
Montaigne
Shakespeare
Hobbes
Milton
Locke
Swift
Rousseau
Austen
Wordsworth
Tocqueville
Dickens
George Eliot
Dostoyevsky
Marx
Nietzsche
Tolstoy
Mann
T.S. Eliot
U.S. Constitution
Federalist Papers
Declaration of Independence
Lincoln, Douglas
Lincoln
MLK Jr.
Hawthorne
Melville
Twain
Faulkner
Bible

Table 8: List of select authors and collaborative works by multiple authors in Bennett's literary canon (1984).

Author
William Shakespeare
Dante Alighieri
Geoffrey Chaucer
Miguel de Cervantes
Michel de Montaigne
Molière
John Milton
Samuel Johnson
Johann Wolfgang von Goethe
William Wordsworth
Jane Austen
Walter Scott
Emily Dickinson
Charles Dickens
George Eliot
Leo Tolstoy
Henrik Ibsen
Sigmund Freud
Marcel Proust
James Joyce
Virginia Woolf
Franz Kafka
Jorge Luis Borges
Pablo Neruda
Fernando Pessoa
Samuel Beckett

Table 9: List of select authors in Bloom's literary canon. (1994)

B Methodology

B.1 Author Title Generation Parameters and Postprocessing

The parameters for generation were kept consistent for each run. Most importantly, each run used greedy sampling (or a temperature of 0) which ensured the most likely result (highest probability) of the LLM's learned token generation distribution. Evaluating results on higher temperatures (leading to more diverse and random outputs) would be a natural follow-up to this study. The other major parameter was that of maximum output tokens which was set determined by the number of titles asked to be generated in the prompt (Table 10).

Because the models output were inconsistent in structure, GPT 4o was prompted to convert the unstructured text output into JSON format. The post-processing prompt and model parameters are given in Table 11. While a few errors occurred in matching the correct author to title, these errors were minimal and fixed manually.

N	Max Tokens
5	250
10	500
20	700
50	1000
100	2000

Table 10: Max token parameters per prompt style.

B.2 Title and Author JSON Postprocessing

System Prompt	For each line, extract the author, title, and year published (if available).
User Prompt	[Previous Output]
Temperature	0

Table 11: Title and author JSON postprocessing prompts and parameters.

B.3 Demographic Generation Parameters and Postprocessing

We used greedy sampling to limit the models to produce the output that is most probable. Additionally, we limited the output to 100 tokens.

String matching for relevant words was used to create flags for each of the three demographic categories: gender, race/ethnicity, and nationality. If flags appear contradictory (such as in the case of gender) or unexpected, a manual check of the output was conducted and the flags were corrected if needed. The algorithms used are provided below.

```
def create_gender_flags(text):
    text = text.strip()
    text = text.replace("\n", " ")
    text = text.replace(".", " ")
    female = 0
    male = 0
    non_binary = 0

    if " male " in text:
        male += 1
    if "**male**" in text:
        male += 1
    if " he " in text:
        male += 1
    if " man " in text:
        male += 1
```

```python
    if "**man**" in text:
        male += 1
    if "he/him" in text:
        male += 1

    if " female " in text:
        female += 1
    if " she " in text:
        female += 1
    if " woman " in text:
        female += 1
    if "she/her" in text:
        female += 1
    if "**female**" in text:
        female += 1
    if "**woman**" in text:
        female += 1

    if "non-binary" in text or "nonbinary" in text or "non binary" in text:
        non_binary += 1
    if "they/them" in text:
        non_binary += 1

    return female, male, non_binary
```

Listing 1: Gender Postprocessing

```python
race_ethnicities = {
    "Asian": ["Asian", "Japanese", "Chinese", "Korean", "Taiwanese", "Indian", "Pakistani",   "Bangladeshi", "Bengali", "Singaporean", "Sri Lankan", "Vietnamese", "Daur Mongol", "Mongolian", "Filipino", "Filipina", "Sri Lankan", "Sri Lanka", "Punjabi", "South Asian"],
    "Black": ["Black", "black", "African", "African American", "African-American", "Afro", "afro", "Nigerian", "Nigeria", "Ghanaian", "Ghana", "Kenyan", "Kenya", "Zanzibari", "Zanzibar", "Cameroonian", "Cameroon", "Jamaican", "Jamaica", "Senegalese", "Senegal", "Haiti", "Haitian", "Congo", "Congolese", "Sudan", "Sudanese", "Zimbabwean", "Zimbabwe", "Somali", "Somalian", "Somali", "Barbadian", "Barbados"],
    "Latino": ["Latino", "Latina", "Latine", "Latinx", "Hispanic", "Mexico", "Mexican", "Colombia", "Colombian", "Chile", "Chilean", "Ecuador", "Ecuadorian", "Argentina", "Argentinian", "Argentine", "Dominican", "Cuba", "Cuban", "Peru", "Peruvian", "Puerto Rica", "Puerto Rican", "Brazil", "Brazilian", "Nicaragua", "Nicaraguan"],
    "Middle Eastern or North African": ["Middle Eastern", "North African", "Arab", "Afghani", "Morocco", "Afghanistan", "Palestinian", "Palestine", "Moroccan", "Numidian", "Iranian", "Iran", "Berber", "Lebanan", "Lebanese", "Oman", "Omani", "Egypt", "Egyptian", "Algeria", "Algerian", "Bahrain", "Bahraini", "Iraq", "Iraqi", "Kuwait", "Kuwaiti", "Libya", "Libyan", "Qatar", "Qatari", "Saudia Arabia", "Saudia Arabian", "Tunisia", "Tunisian", "UAE", "Emirati", "Yemen", "Yemeni", "Jordan", "Jordanian"],
    "Native American": ["Native American", "Indian American", "indigenous", "Indigenous", "Lakota", "Blackfeet", "Spokane", "Cheynee", "Arapaho", "Ojibwe", "M\u0x00E9tis", "Metis", "Anishinaabe"],
    "Pacific Islander": ["Pacific Islander", "Maori", "M\u0x0101ori"],
    "White": ["White", "white", "European", "Caucasian"]}

def create_race_ethnicity_flags(row):
    text = row["output"]
    author = row["author"]
    text = text.strip()
    text = text.replace(".", " ")
    text = text.replace(",", " ")
    text = text.replace("\n", " ")

    race_ethnicity_flags = {"race_pred_" + key: 0 for key in race_ethnicity_flags.keys()}
    race_ethnicity_flags["race_pred_Not Mentioned"] = 0
    race_ethnicity_flags["author"] = author
    race_ethnicity_flags["output"] = text
    mention = 0

    for key in races:
        for valid_word in race_ethnicities[key]:
            if valid_word + " " in text or valid_word + "-" in text or "**" + valid_word + "**" in text:
                mention = 1
                race_ethnicity_flags["race_pred_" + key] += 1

    if mention == 0:
        race_ethnicity_flags["race_pred_Not Mentioned"] = 1

    return race_ethnicity_flags
```

Listing 2: Race/Ethnicity Postprocessing

C LitAuthorDemoDB

Author Table	Book Table
Author ID	Book ID
First Name	Author ID
Last Name	Book Title
Middle Name	Author Full Name
Known Aliases	Year Published
Gender	
Race/Ethnicity	
Nationality	
Notes	

Table 12: Features of the author and book tables for LitAuthorDemoDB.

D Results

					Western	Global	
Model	**None**	**Fiction**	**Classic**	**Literary Canon**	**Literary Canon**	**Literary Canon**	**Total**
GPT 4o	268	206	227	217	197	217	466
Llama 3.1 8B	256	221	211	162	190	156	541
Llama 3.3 70B	377	317	324	294	333	314	745
Gemini 1.5 Pro	367	330	305	282	266	320	720
Total	695	603	622	528	573	571	1346

Table 13: Unique authors by model and prompt description.

					Western	Global	
Model	**None**	**Fiction**	**Classic**	**Literary Canon**	**Literary Canon**	**Literary Canon**	**Total**
GPT 4o	318	274	318	296	306	274	711
Llama 3.1 8B	305	263	267	248	317	199	841
Llama 3.3 70B	518	433	444	422	448	396	1108
Gemini 1.5 Pro	478	449	402	368	364	399	1035
Total	1027	909	891	812	899	786	2238

Table 14: Unique titles by model and prompt description.

nationality	n	p
American	682	0.525
British	259	0.2
English	141	0.109
French	47	0.036
Canadian	36	0.028
Irish	35	0.027
Australian	23	0.018
German	22	0.017
Greek	20	0.015
Italian	19	0.015
Japanese	17	0.013
Russian	17	0.013
Scottish	17	0.013
Chinese	16	0.012
Indian	15	0.012
Roman	13	0.01
Nigerian	13	0.01
Austrian	9	0.007
Mexican	7	0.005
Argentinian	7	0.005
New Zealand	6	0.005
Iranian	6	0.005
Swedish	6	0.005
Dutch	6	0.005
South African	5	0.004
Vietnamese	5	0.004
Swiss	4	0.003
Spanish	4	0.003
Polish	4	0.003
Malaysian	4	0.003
Colombian	3	0.002
Welsh	3	0.002
Sri Lankan	3	0.002
Persian	3	0.002
Taiwanese	3	0.002
Ghanaian	2	0.002
Czech	2	0.002
Chilean	2	0.002
Jamaican	2	0.002
Pakistani	2	0.002
South Korean	2	0.002
Zimbabwean	2	0.002
Peruvian	2	0.002
Lebanese	2	0.002
Portuguese	2	0.002
Turkish	2	0.002
Romanian	2	0.002
Palestinian	2	0.002
Israeli	2	0.002
Danish	2	0.002
Hungarian	2	0.002
Norwegian	2	0.002
Korean	2	0.002
Barbadian	1	0.001
Cyproit	1	0.001
Singaporean	1	0.001
Mesopotamian	1	0.001
Congolese	1	0.001
Egyptian	1	0.001
Numidian	1	0.001
Iraqi	1	0.001
Khwarezmian	1	0.001
Ecuadorian	1	0.001
Icelandic	1	0.001
Cameroonian	1	0.001
Albanian	1	0.001
Norman	1	0.001
Nicaraguan	1	0.001
Ukranian	1	0.001
Haitian	1	0.001
Unknown	1	0.001
Brazilian	1	0.001
Berber	1	0.001
Sudanese	1	0.001
Bahamian	1	0.001
Senegalese	1	0.001
Omani	1	0.001
Finnish	1	0.001
Moroccan	1	0.001

Table 15: Proportion of authors by nationality.