RESEARCH

Mapping Language Literacy At Scale: A Case Study on Facebook

Yu-Ru Lin[1*], Shaomei Wu[2] and Winter Mason[3]

[*]Correspondence: yurulin@pitt.edu
[1]School of Computing and Information, University of Pittsburgh, Pittsburgh, PA, USA
Full list of author information is available at the end of the article

Abstract

Literacy is one of the most fundamental skills for people to access and navigate today's digital environment. This work systematically studies the language literacy skills of online populations for more than 160 countries and regions across the world, including many low-resourced countries where official literacy data are particularly sparse. Leveraging public data on Facebook, we develop a population-level literacy estimate for the online population that is based on aggregated and de-identified public posts written by adult Facebook users globally, significantly improving both the coverage and resolution of existing literacy tracking data. We found that, on Facebook, women collectively show higher language literacy than men in many countries, but substantial gaps remain in Africa and Asia. Further, our analysis reveals a considerable regional gap within a country that is associated with multiple socio-technical inequalities, suggesting an "inequality paradox" – where the online language skill disparity interacts with offline socioeconomic inequalities in complex ways. These findings have implications for global women's empowerment and socioeconomic inequalities.

Keywords: global literacy; global inequalities; social media demography; information accessibility; cross-language measurement

1 Introduction

Literacy, the ability to comprehend and produce textual information, is known as the foundation for many important personal and social functions. For individuals, the lack of literacy skills is associated with reduced access to education [1, 2, 3], employment [1, 4, 2, 5, 6], social benefits [7], as well as poorer health outcomes [8, 9] and lower civic engagement [1, 4, 9, 10]. Collectively, literacy is considered a prerequisite for democracy and socioeconomic development [6, 10].

Despite a substantial increase in global literacy rates over recent decades, there were still 750 million adults – two-thirds of whom were women – remaining illiterate in 2016 [11]. The rise of digital communication technology has brought new challenges to those with limited literacy skills: as more and more public, professional, and social communications shift to the digital, text-mediated environment, a lack of literacy skills can not only exclude people from the information and resources available online but also expose them to greater (mis)informational vulnerability and harms [12, 13]. With most existing literacy programs and research focusing on school children and educational settings, we see a significant gap in our understanding of *literacy practices and challenges in the digital environment*. In this study, we take a data-driven approach, leveraging the data available on Facebook – the most

widely adopted social media platform with a third of the world's population using it regularly [14] - to obtain a representative and up-to-date sample of literacy activities (e.g. reading and writing textual content) by the global online population.

This study systemically examines the *language literacy skills of online populations* (henceforth called *online language literacy*) for more than 160 countries and regions around the globe. We introduce a new population-level measure called *online language literacy estimate* (`OLLE`) that is based on aggregated and de-identified written content posted publicly on Facebook. Thanks to the reach of Facebook to hundreds of millions of active users from low-resourced regions such as Africa, Latin American, and South East Asia [15], our measure is able to estimate and track population-level language literacy at an unprecedented level of coverage, resolution, and timeliness comparing to traditional literacy assessment methods [16], while achieving an overall strong correlation with available official data. With `OLLE` calculated for different gender, country, and regional groups across the world, we capture the disparities in online literacy across broad geographical areas and explore gender and regional literacy gaps under a diverse set of societal contexts. Our results not only quantify the association between online language literacy gaps and offline inequality metrics, but also uncover the complex interaction between literacy, Internet adoption, and civic participation for women. In summary, the main contributions of our work are:

- We develop a global online language literacy estimate (`OLLE`) using Facebook data from over 160 countries in 12 languages.
- We evaluate our measure with existing offline population literacy benchmarks, showing a strong correlation and broader coverage than current official data.
- We demonstrate how the online language literacy measure can be used to track gender and regional literacy gaps and unpack the complex societal context around literacy and literacy development.

The rest of the paper is structured as follows: Sec. 2 offers a literature review of related work to contextualize our study. Sec. 3 describes our methodology and the dataset used for developing the online language literacy estimate (`OLLE`). Sec. 4 validates the resulting `OLLE`'s with existing literacy assessment data and presents an overview of online language literacy skills across the world. We also share a few applications of `OLLE` in studying and understanding regional and gender inequalities globally. Sec. 5.1 discusses the implications and limitations of this study, and concludes our work.

2 Background and Related Work

2.1 Population Literacy Assessment

Recognizing the importance of literacy in reducing poverty and expanding lifelong opportunities, the United Nations has included *literacy* as part of its Sustainable Development Goals (Goal Target 4.6) [12, 17]. However, tracking population-level literacy development for different demographics globally remains challenging, with most existing datasets incomplete, dated, or costly to obtain [18].

Worldwide, the United Nations Educational, Scientific and Cultural Organization (UNESCO) has been tracking country-level literacy rates for major demographics such as youth, adults, men, and women [19]. However, their data is based on self-declaration of reading and writing skills, often collected by asking the head of the

household to answer questions like: "*Can you (and others in your home) read and write a simple sentence?*" As a result, the data may overstate actual skills and not capture any notion of functional literacy [18]. Even after adding a simple test of reading skills in the data collection process, the results only group people into three big categories – illiterate, functional literate, and literate – and cannot measure literacy on a continuum. Despite issues in the UNESCO data, they are still a major reference point, especially for developing countries and regions where the government infrastructures for census and population surveys are scarce.

In the developed world, many countries have invested significant efforts to develop and implement modern literacy assessments that capture population literacy skills beyond a simple *literacy-illiteracy dichotomy*. In the US, the National Adult Literacy Survey (NALS) has been funded by the federal government in 1992 and 2003 [1, 4]. Internationally, there have been coordinated efforts to assess adult literacy skills through programs such as the Program for International Assessment of Adult Competencies (PIAAC), involving 39 countries and regions since its inception in 2012 [20]. While these assessments provide more granular and contextualized literacy skill measures, they are expensive to administrate and hard to scale: both NALS and PIAAC were conducted once per decade, requiring 8 to 10 months to conduct the surveys and interviews, and a few years to compile the results [1, 4, 20].

As a result, the Global Alliance to Monitor Learning has recently made the call to develop "efficient", and "light" methodologies to gather nuanced, standardized data that allows for cross-national tracking and comparison [18]. This work directly responded to this call, by proposing a data-driven method that leverages social media data to estimate the literacy skills of diverse geographic and demographic populations in a cost-effective way with unprecedented coverage. Although our data were collected from only one platform – Facebook, its high penetration in many parts of the world allows our method to capture the literacy skills of the entire population, especially populations with high Internet adoption.

2.2 Digital Literacy

Although closely related to *digital literacy* – the ability to operate and communicate through digital technology [21], language literacy is composed of fundamental language skills such as reading, writing, and numeracy, that are often a prerequisite for digital literacy [13, 22, 23]. In fact, research has shown that the lack of language literacy skills is a top barrier to Internet access and technology adoption [13, 22, 23].

As human society enter an increasingly technological and informational-rich age, modern literacy assessment programs such as PIAAC also include the assessment of "problem-solving skills in technology-rich environment", showing several demographic differences and similarities in literacy and digital literacy proficiency in the developed countries [9]. For example, while the gender gap in favor of men was observed with digital literacy skills, there is a very small or non-existent gender difference in literacy skills. Similarly, the age gap in favor of young people was more observed with digital literacy than with literacy [9]. The results from PIAAC also showed a significant interaction effect between gender, age, and socio-economic backgrounds on literacy and digital literacy [9], inspiring us to explore similar trends for the broader global population covered by this research.

With over 27,000 new Internet users every hour and many of them from traditionally low-resourced regions [24], this work measures and characterizes the language literacy skills of the population that is already online - as captured on Facebook, laying the foundation for future research on more contextualized literacies such as digital literacy and information literacy.

2.3 Literacy and Social Media

Most studies of literacy in the social media context focus on youth and their literacy practice. For example, many studies documented the young social media users' practice of "remixing" - creating, uploading, selecting, copy-pasting, combining, and co-producing content in their profiles and timelines, noting a new literacy practice that is more collaborative, dynamic, and multi-model than traditional, print literacy [25, 26, 27, 28, 29, 30]. As a result, the task of "reading" has changed significantly, becoming more technically simple yet socially complex while [31, 25].

Numerous research examined the relationship between social media and literacy development, especially, for socially disadvantaged groups and English Language Learners (ELLs). Most of these works supported the benefits of social media in literacy development. For example, [32] found that reading and writing blogs enhanced the confidence in writing for young people in the UK; [33] documented the use of Facebook is positively correlated with improved English skills for college students in Indonesia; and [34]) argued that social media technologies can support ELLs to develop valuable print literacy, based on longitudinal ethnographic studies of adolescent ELLs literate and social activities around online fandom communities. However, some research also suggested that social media use can negatively impact the reading culture and academic performance of students [35].

As misinformation on social media became a public concern [36, 37], some recent work highlighted the importance of literacy skills in the social media age. For example, data from the PISA 2018 reading assessment showed that less than 10% of the 15-year-old students in OECD countries had the reading proficiency level to distinguish facts from opinions, which could significantly impact their abilities to assess the quality and credibility of information spread on social media [38]. While an increasing amount of attention has been devoted to developing the digital and media literacy of online populations, this work underlines the prevalence of language literacy challenges and calls for future research in understanding the scale and impact of misinformation on low literacy populations.

3 Methods

3.1 Facebook dataset

To obtain a written sample of online populations worldwide, we collect public posts written in any of the 12 chosen languages created by Facebook users who are at least 18 years old and active during a 30-day period between April 20 and May 20, 2020. To ensure that the collected posts represent the writings of individual Facebook users, we exclude posts made by pages, organizational accounts, and public profiles. We also exclude posts that did not contain any text or text that was shorter than 2 characters or longer than 1000 characters, as well as posts that contained URLs as these are more likely to be copied and pasted from other sources rather than

composed by users. After pre-processing the text, the median length of posts in our dataset is 18 characters, the average is 51 characters, and the 95th percentile is 187 characters.

All the analysis and statistics were computed over a populational aggregate, based on self-reported attributes such as country and gender. `OLLE` were only calculated for groups with at least 1,000 active users, to minimize the risk of user de-anonymization. The original text, together with the intermediate results, were dropped after populational level `OLLE` and other statics were calculated.

More details about Facebook public post data collection and use can be found in *Supplemental Material*: S1.3.

3.2 Estimating online language literacy

We use the vast amount of text produced by Facebook users to quantify populations' online language literacy skills. This method relies on two assumptions: (i) literacy across populations may be measured collectively by aggregating the observed data within country or regional borders, and (ii) vocabulary usage patterns observed in corpora of written texts can be used as a proxy for language practice in a digital environment. The first assumption comes from prior observations that literacy skill is locally clustered, largely determined by the socioeconomic status of the local community and by the abundance of public resources such as public school systems and libraries [2]. This also alleviates the need to analyze individual-level data, which are harder to de-identify, and often too sparse for reliable estimates. The second assumption is motivated by the relationship between literacy and vocabulary knowledge found in the education domain [39, 40, 41, 42, 43]. In practice, vocabulary size was often employed as a proxy for literacy skill, as reading comprehension cannot be achieved unless the reader knows 95% of the words in the text [44], and a certain vocabulary size is required for unassisted text comprehension [45]. Vocabulary knowledge of words at various frequency levels has been used to measure total vocabulary size. For example, the Vocabulary Size Test (VST) tests a learner's knowledge of 140 or 200 words, with 10 words sampled from each 1000 word in frequency levels based on British National Corpus [46].

Analogous to the VST method, we measure the aggregated use of lower-frequency words ("LoFF words") - secondary vocabulary words outside the high-frequency everyday lexicons - in public Facebook posts as a proxy for online populations' literacy skills in a given language. Intentionally designed as a population-level, aggregated measure, our approach does not collect any personally identifiable information or any personal/private content (see *Supplemental Material*: S1.3). For maximal geographical coverage, we pick the twelve (12) most widely used languages (in terms of countries) and algorithmically define a set of lower-frequency words for each language. The 12 selected languages are: Arabic (ar), German (de), English (en), Spanish (es), French (fr), Italian (it), Malay (ms), Dutch (nl), Portuguese (pt), Russian (ru), Turkish (tr), and Chinese (zh). The next section will detail our method to determine the sets of LoFF words based on a multi-lingual reference corpus.

3.3 Detecting LoFF words from a reference language corpus

While the VST method relies on the British National Corpus (BNC) for baseline word-frequencies, we use the fastText unigram data [47] as our reference corpus.

Comparing to other popular language corpora such as BNC and Google Books Ngram [48], fastText has two major advantages: (i) coverage: The distributed fastText data currently supports 157 languages and covers all 12 languages considered in our study. (ii) up-to-date representation of vocabularies used by online population: fastText is based on texts collected from Wikipedia and Common Crawl [1], containing petabytes of web page data collected since 2008.

Using fastText data, we are able to retrieve up 200,000 most frequent words in each language as the candidates for LoFF words in that language. Understanding that the exact range of LoFF words may vary depending on the language, we first try to understand the use of these top 200,000 words by Facebook users through language-specific "word popularity curves": a scatter plot where the x-axis represents the rank of a word by its frequency (0 for the most frequent), and the y-axis represents "word popularity", as measured by the percentage of unique users in the language population who have used that word in our data. Fig. S1 (*Supplemental Material*) shows the word popularity curves for the 12 chosen languages in our study.

As illustrated in Fig. S1, the word popularity curves first decline sharply for the most frequent words before settling into a relatively flat region. The transitional area - the "elbow" (or "knee") region of the smoothed word popularity curve - corresponds to the words that are neither too popular nor too unpopular, thus ideally covering the set of LoFF words for that language. Mathematically, the curvature is a mathematical measure of how much a function differs from a straight line [49, 50], and the elbow areas start at the point of maximum curvature in a popularity curve. Estimating the knee/elbow point for a continuous function is straightforward since the curvature is well-defined for continuous functions; however, it is a challenging task for discrete data. We leverage the "Kneedle" detection [49], an efficient algorithm that can efficiently detect knee points in discrete data, and the standard maximum curvature approach on a smoothed function learned from the discrete points. We found this hybrid approach is more robust to rescaling and small fluctuations in our data. Details of this elbow detection is given in *Supplemental Material*: S1.2.

Fig. 1 A-C highlight the elbow range detected from the word popularity curve for each of the three most used languages (English, Spanish, and Arabic). Words falling into the elbow range are defined as the "LoFF words" in the language. Fig. S1 shows the detected elbow ranges for all 12 languages.

3.4 Calibration for language use and bias

After determining the LoFF words per language, we can calculate the relative frequency of LoFF words, among all the public text posted by Facebook users from a given country in a given language. We denote the relative frequency of LoFF words as $\bar{w}_{c,l}$, with l representing the 12 languages considered in this study and c representing the 167 countries whose official or dominant languages are among these languages.

While $\bar{w}_{c,l}$ makes it possible to track online language literacy for multiple languages in parallel, we decide to use one representative language per country in all our analyses for simplicity. For most countries, the official language is chosen as the

[1] https://commoncrawl.org/

representative language. For countries with no or multiple official languages, we use the language that is used by most users in that country. For example, for India, a country that uses both Hindi and English as official languages, we use English as the representative language for India since English was used by the largest number of Facebook users in India based on our data. Fig. S2 (*Supplemental Material*) shows the number of countries broken down by dominant language.

As validation, we group countries by the representative language and compare the value of $\bar{w}_c$ with the official literacy rate data for each country. After excluding countries where the official literacy rate is not available or where the Internet penetration is lower than 25% percent, the three most widely used languages (in terms of the number of countries covered) in our data are English (41 countries), Spanish (19 countries), and Arabic (15 countries). Fig. 1 D-F shows the relationship between the officially reported literacy rates and $\bar{w}_c$, for English, Spanish, and Arabic, respectively: each dot represents a country, with x value corresponding to the rank-based quantile normalization[51] of $\bar{w}_c$, and y value corresponding to the rank-based quantile normalization of its official literacy rate. The shadow areas in Fig. 1 D-F visualize Spearman's rank correlation coefficient between the two variables, with 95% CIs[2]. As seen in Fig. 1 D-F, all three sets of countries exhibit strong positive correlations, though the Spanish-speaking countries show a larger variance than countries mostly using English or Arabic. The positive correlations between $\bar{w}_c$ and the reported literacy rate in the three top languages suggest the efficacy of using the relative LoFF word count as a measure of literacy across the different most-used languages. The correlations for other languages are not reported because none of the remaining languages cover more than 7 countries.

To generate a global online literacy estimate that is directly comparable across languages, we also calibrate $\bar{w}_c$ for all countries using the official literacy data collected by UNESCO [19]. Given that the two measures have different distributions, both $\bar{w}_c$ and the official literacy data were transformed to better fit a normal distribution. A rank-based ordered quantile normalization transformation [51] was used for both $\bar{w}_c$ and the literacy rates, where the transformation $g(\cdot)$ is given by $g(x_i) = \Phi^{-1}\left(\frac{r_i - 0.5}{n}\right)$ where Φ refers to the standard normal CDF, x_i is a continuous measurement observed for each object i, r_i refers to the sample rank of i when the measurements are placed in ascending order, and n refers to the number of observations. In the case of new values that fall outside the observed domain of x, we adopt the standard procedure and use generalized linear models to estimate the ranks beyond the bounds of the original domain of x.

A linear fixed effect model is employed to quantify the systematic differences across languages *in relation to the offline literacy rates*. Let $l^{(i)}$ be the representative language for a (country) population i, the calibrated estimate, denoted as $\hat{w}_i$, is given as $\hat{w}_i \propto y_{il}$, where

$$y_{il} = \beta x_{il} + \alpha_l + \epsilon_{il},$$

[2]Without specification otherwise, the confidence interval for all the correlation coefficients are produced using a nonparametric bootstrap procedure based on the percentile method (with 1000 bootstrap replicates).

where $x_{il} = g(\bar{w}_i)$ is the direct literacy estimate of population i, α_l is the language-specific effect, ϵ_{il} is the idiosyncratic error term with $\epsilon_{il} \sim \mathcal{N}(0, \hat{\sigma}_\epsilon^2)$, and β is the parameter. After β is learned, **the global online language literacy estimate (`OLLE`) is the calibrated estimate** $\hat{w}_i$, obtained by rescaling y_{il} between 0 and 1. This rescaling step is to make the `OLLE` value more interpretable and to facilitate the comparison across populations and subpopulations by a single index.

Since the information about the official literacy rates has been used in the calibration, to get an unbiased evaluation, we use the leave-one-out procedure to obtain an out-of-sample evaluation – for each i, the calibrated estimate was generated by using all countries other than i in the model (the estimated parameters can be found in Model (a) in Table S2). In other words, the language calibration is calculated from the residualized mean aggregated over the rest of the countries using the same language as i.

While the official literacy rates data has served an important role in our study to validate and calibrate `OLLE`, they are not suitable as target variables for building a predictive model for online language literacy, for a few reasons: 1) conceptually, we want to distinguish online population literacy and general population literacy in this study; 2) technically, the official literacy rates data were collected for different countries at different time, with methodological variances over the years, introducing extra noise and latent variables for a robust supervised model.

3.5 Country-level socioeconomic covariates

To understand the relationship between literacy and other social factors, we collect information about countries' socioeconomic status and technical development from multiple data sources. Tables S6 and S8 list all variables used in our study, with definitions and the sources where the variables are gathered. The first-order correlations among these variables are provided in Table S6. Due to the heterogeneous distributions across variables, we report correlations using Spearman's rank correlation coefficient unless otherwise stated.

Inspired by previous research showing the effect of a country's income, Internet penetration, and other gender inequality measure on the digital gender gap [52], we study the relationship between social factors and online literacy gaps through regression analysis with the gender or regional differences in `OLLE` as the dependent variable and country-level variables as the independent variables. All the regression models presented in this work are based on standard OLS estimates, where all variables are first separately transformed to better fit a normal distribution. For example, the income variable was transformed logarithmically. Considering the geographical clustering of many of the socioeconomic variables, in each of the regression analyses, we provide models with and without controls for geographical groups. The control for geographical groups signifies whether the pattern observed in our study is a global phenomenon or particular to certain areas. For example, when predicting the gender gap in `OLLE`, the coefficient estimates are consistent between the models with and without geographical information. We show in *Supplemental Material* Tables S9-S11 the detailed estimates of regression models and their comparisons. The consistent coefficient estimates are also found in predicting regional disparity (*Supplemental Material* Tables S12-S13).

4 Results

The results are two-fold. We first validate of our literacy estimates with existing official data and report the global state of online language literacy across 167 countries. Next, we consider within-country demographic segments such as gender and regions to benchmark the differences in online language literacy across subpopulations, and examine the socioeconomic factors that explain those differences.

4.1 Online Language Literacy Worldwide

4.1.1 Significant agreement between OLLE*'s and official literacy data*

Following the methodology detailed in Section 3, we generate OLLE, the calibrated online language literacy estimates, in 167 countries or regions whose representative languages are among the 12 selected languages and have at least a thousand adult Facebook users who posted publicly in the representative language during our data collection period. To ensure we obtain a sufficient sample size of the population in each country, we leave out five countries—China, Iran, Russia, Kazakhstan, and Turkmenistan—where Facebook use is curbed by the countries' government regulations or other policy challenges. We show the country and world population coverage breaking down by languages in *Supplemental Material* Fig S2. The raw and calibrated values of online language literacy estimates for all 167 countries can be found in *Supplemental Material* Table S4.

Fig 1 illustrates the key steps of our methodology, as well as the correlation between OLLE's and official literacy rate data in 1G. As visualized in Fig. 1G, we find a strong and positive correlation between OLLE with the reported literacy rates (Spearman's rank correlation $\rho = 0.78$, 95% CI [0.69, 0.84], $p < 0.001$ based on out-of-sample evaluation). Similar results are found when validating our estimates with global educational attainment statistics: OLLE is highly correlated with a country's average schooling years ($\rho = 0.78$, 95% CI [0.65, 0.87], $p < 0.001$; see details in *Supplemental Material* Fig. S3). These findings indicate that our estimates do reflect populations' literacy skills and can be used as a reliable proxy for official literacy and educational attainment statistics when such data are unavailable or outdated.

4.1.2 Understand global online literacy inequalities through OLLE

One direct application of OLLE is to track the state of literacy for online populations across the globe. Mapping OLLE's by country in Fig. 2B, we see significant inequalities in online language literacy skills across geographical regions, with the "global south" countries collectively lagging behind in terms of online population literacy skills. Fig. 2A summarizes the aggregated statistics for seven geographical groups and benchmarks the online literacy gaps across regions. The bottom 10% of countries with the lowest OLLE's are primarily located in Sub-Saharan Africa (13 countries), plus one in Latin America & the Caribbean (Haiti), and one in Northern Africa & Western Asia (Algeria). While our result is consistent with the geographical patterns observed in official literacy rate data [11], it also highlights the persistence of literacy gaps across offline and online populations, calling out for additional literacy support for a substantial percentage of the online population in today's digital environment.

On the other end of the spectrum, the top-ranked countries in terms of OLLE are all located in the Europe & North American region, as well as Oceania, with the

top 3 countries being Belarus, Ukraine, and San Marino. While this result is again largely consistent with the official literacy rate data by UNESCO, it also suggests potential biases introduced by language-based calibration. For example, countries with Russian as the representative language (e.g. Belarus, Ukraine) could get an extra boost during calibration due to the overall high literacy rates in Russian-speaking countries reported in the official data.

4.2 Gender Difference in Online Language Literacy

Although the gender gap in literacy has been shrinking globally in recent decades, women are still facing obstacles when accessing school and the Internet [11, 53]. As a result, serious male-favoring gender gaps in literacy skills still persist in the Middle East and North Africa, South Asia, and Sub-Saharan Africa regions [54]. An earlier study showed that, in low-income countries, female Internet penetration is 24% lower than that for males [52]. To track the gender literacy gap in the online population, we calculate the standardized difference in `OLLE` between female and male Facebook users in each country. This measure captures both the direction and the size of the gender gap in online language literacy, where a positive value indicates a female-favoring gap, and vice versa. To ensure a sufficient sample size of the male and female subpopulations, we drop countries (17 out of 167) with fewer than a thousand adult users in either gender group in our dataset for this analysis. While the gender digital divide has generally referred to the gaps in access and use of digital technology, our measure calls for attention to the disparity in language skills between women and men who are already online. Understanding that gender is non-binary, we only present the female-male gender gap here for two reasons: (a) it enables us to correlate our estimates with existing gender gap data; (b) in our dataset, we do not have sufficient data from users with self-reported non-binary gender information to deliver a reliable estimation for this sub-population.

4.2.1 Women collectively scored higher than men on Facebook in most countries, with substantial male-favoring gaps in two regions

Fig. 3A shows the gender differences in `OLLE` captured in our data. Among the 160 countries where the gender gap in `OLLE` is calculated, 69 countries (43.1%) have significant female-favoring gaps and 54 countries (33.8%) have significant male-favoring gaps. The remaining 37 countries do not have a significant gender gap in `OLLE`. The significance is determined based on whether a country's male-favoring gap (or female-favoring gap) falls above the 95% confidence limits of the expected male-favoring gaps (or female-favoring gaps). Overall, we observe more countries having female-favoring gaps in our measure, suggesting on average higher language literacy skills for women than men among today's online population.

Fig. 3A highlights countries with the most and least substantial female-favoring gaps. As observed in Fig. 3A, almost all of the countries with the world's largest advanced economies (the G7) have a significant female-favoring gap, with Italy the only exception. This finding is generally consistent with the data collected in recent PISA tests, which showed that girls outperformed boys in reading in all the OCED countries and regions [55]. Despite progress, gender-based inequalities are still pervasive in today's society. To summarize the global state of online literacy

difference between male and female subpopulations, a world map of OLLE gender gap is provided in *Supplemental Material* Fig. S5. Notably, while most regions on average appear to have female-favoring gaps, two regions (Sub-Saharan Africa and Northern Africa & Western Asia) still show substantial gaps favoring men. In Sub-Saharan Africa, there are 22 countries with male-favoring gaps, compared to 8 with female-favoring gaps; in Northern Africa & Western Asia, there are 15 countries with male-favoring gaps, compared to only two countries with female-favoring gaps. The regional patterns here are generally consistent with what was reported by UNESCO in the official literacy rate data [11]. Our measure, however, characterizes more countries with female-favoring literacy gaps than the official data, indicating a potential populational difference between women on Facebook and the general female population in a country. We will further explore the relationship between populational factors and the gender gap in OLLE in the next subsection.

4.2.2 Understand the societal context for gender online literacy gap

We first compare the observed gender gap in OLLE with other country-level measures such as overall OLLE, income per capita, Gini index, average education attainment, and Internet penetration rate. As shown in Fig. 3 (B,C,D) the OLLE gender gaps are positively correlated with the countries' OLLE (Spearman's rank correlation $\rho = 0.59$, 95% CI [0.47,0.68], $p < 0.001$), overall education ($\rho = 0.59$, 95% CI [0.44,0.71], $p < 0.001$), and Internet penetration ($\rho = 0.30$, 95% CI [0.14, 0.46], $p < 0.001$), suggesting that women are disproportionally disadvantaged in low-resourced, low-literacy countries. Fig. 3 (E,F,G) show that countries' OLLE gender gaps significantly correlate with other gender parity measures, including a positive association with the female-male difference in offline literacy rates ($\rho = 0.43$, 95% CI [0.26, 0.58], $p < 0.001$), women's civic participation ($\rho = 0.48$, 95% CI [0.3, 0.62], $p < 0.001$), and a negative association with the countries' Gender Inequality Index or GII ($\rho = -0.4$, 95% CI [-0.57, -0.23], $p < 0.001$). The GII reflects how women are disadvantaged in multiple dimensions of human development and thus a negative association is expected [56]. Interestingly, among all gender parity or empowerment measurements, women's civic participation – the extent to which women have the ability to express themselves and to participate in civil society [57] – appears to have the strongest association with the OLLE gender gap. This could suggest that the offline structural barriers to women's civic participation are strongly associated with their literacy relative to men in the online space.

To better understand the societal context for online literacy gender gap globally, we further examine the relationship between country-level variables and the observed gender gap in OLLE through multiple regression Ordinary Least Squares (OLS) model. Given that many of the country-level variables are highly correlated (see *Supplemental Material* Table S6), we only include the most relevant variables in this analysis. Fig. 4A summarizes the estimated effect of these variables, where the effect of geographical grouping is further detailed in *Supplemental Material* (see Tables S9–S11). Based on the OLS estimation, the OLLE gender gap remains significantly and positively associated with overall education status and women's civic participation while controlling for other variables. The overall Internet penetration rate is negatively associated with the OLLE gender gap. This may look counterintuitive. One possible interpretation is that a lower level of Internet penetration rate

excludes groups from lower socioeconomic status to participate in the digital world, and women in those groups also tend to lack opportunities in education and many other developmental aspects. Hence, a lower level of Internet penetration rate ironically serves as an equalizer for the OLLE gender gap. Interestingly, the OLS model also reveals an interaction effect between the overall Internet penetration rate and women's civic participation on the OLLE gender gap. When a country's Internet penetration rate is high, the country's OLLE gap may be either high or low – depending on whether the country has a high level of women's civic participation (Fig. 4B). This could suggest that technological advancements – i.e., the adoption of the Internet – are not necessarily associated with more opportunities or higher skills for women relative to men unless such a relationship appears in a society where women have the chance to actively participate in civic processes.

4.3 Within-Country Regional Disparity in Online Language Literacy

While it is widely acknowledged that the disparity in education resources and technology infrastructure has contributed to the digital divide between developed and developing countries [58, 59], there have been only a few studies that examined the digital disparities within a country – often limited to studying a single country with the digital divide among gender and ethnicity groups [60, 61]. Here we intend to provide insights into the within-country regional digital disparities for a large number of countries across the world. Extending our methodology, we measure the within-country regional disparity in online language literacy by quantifying the variability of regional OLLE's for a given country. More specifically, the variability of regional OLLE's is calculated as the standard deviation of OLLE's aggregated across available regions within a country, thus a larger value indicates a higher variability observed in OLLE's at a sub-national level. We define a region as a sub-national administrative division (self-government or jurisdiction under a country's national laws), such as a state or a province. Countries with less than two regions having a minimum of a thousand active adult Facebook users in our dataset are excluded, which resulted in 119 countries in our analysis. Fig. S6 in *Supplemental Material* provides summary statistics and a map of the within-country regional disparity in OLLE, as well as the representative countries with a relatively high or low level of regional disparity from each geographical group.

4.3.1 Within-country regional disparity in OLLE *is associated with multiple inequality measures*

We examine the societal backdrop for the observed regional disparities in online language literacy. Our particular interest is in the link between the regional disparity in OLLE and countries' resource distribution, such as the inequalities in education and income within a country, as well as its overall education and socio-technical development. Using multiple regression analysis, we find that, after controlling for all other variables, inequality in education and the Internet penetration rate have a strong and positive association with regional disparities in OLLE (Fig. 4C). Not surprisingly, a higher level of overall educational attainment predicts a smaller regional variation in online language literacy skills, which is converse to the effect of inequality in education. The inequality in income, as captured by a country's

Gini index, however, appears to have a negative relationship with the OLLE regional disparity, indicating a greater income inequality is associated with smaller regional OLLE disparity within the country.

4.3.2 Inequality paradox

The interaction between inequalities in education and income is also observed (Fig. 4D), where a greater level of within-country regional disparity in OLLE is predicted for countries with one of the two conditions: either the country has a relatively high level of income and education inequalities, or has a relatively low level of both inequalities. This finding suggests an "*inequality paradox*" – a paradoxical pattern we notice that links the offline socioeconomic inequalities to online language skill disparity in surprising ways. For example, in countries with a higher level of both economic and educational inequalities, access to social media is more likely to be reserved for the more socio-economically advantaged groups, and therefore show a less regional disparity in OLLE (corresponding to the bottom-left corner of Fig. 4D). In contrast, in countries where education inequality is low but economic inequality is high, a higher level of regional disparity in OLLE is observed (corresponding to the bottom-right corner of Fig. 4D). Similar patterns are observed when taking the geographical grouping into account, suggesting that the observed patterns are common across societies (see *Supplemental Material* Tables S12–S13 for more details).

5 Discussion

A data-driven, cross-language, cross-country online literacy estimation Taking advantage of the abundance of user-generated text online, our proposed methodology of measuring online language literacy can be scaled across languages and subpopulations, as long as population-level text corpora are available. OLLE complements the traditional sources and makes it possible to monitor future progress and answer questions such as: whether the online language skills improve faster (slower), or whether the literacy gap is closing (widening), particularly in low-literacy countries. While our current dataset only contains text generated within a 30-day period, further collection of similar data over a longer period of time will offer new insights on the temporal evolution of OLLE's across the world.

Tracking global trend in online language literacy This study reveals the current state of global online language literacy. Based on our study, the estimated global online language literacy has remarkably high correlation with documented country-level literacy rates and educational attainments data (with $\rho = 0.78$ in both correlations; see Fig. 1 and *Supplemental Material* Fig. S3). This finding has two implications. First, given that 86% of the world population are now reportedly literate (i.e., able to read and write) [11], our study suggests the variation in language skills remains *within the literate, online population*. Even though many countries now have more than 95% literate populations, our online literacy map (Fig. 2) has revealed the nuanced differences among the online population's language skills in these countries. Second, beyond a few options in assessing a country's digital advancement, such as the Internet penetration measure, OLLE's robust correlation with offline literacy and educational data make it a more relevant alternative for tracking the *outcome* of a country's access to education and resources for global literacy development.

Women's empowerment, social inequalities, and online language literacy disparities In our study, we show that the gender difference in OLLE is significantly correlated with various offline gender parity metrics, including gender gaps in literacy rate, education, GII (which also considers the economic standing across gender groups), and women's civic participation index. This suggests that a country's offline gender equity progress is crucially relevant to how literate populations across genders participate online. In contrast to existing studies that exposed the well-known correlation between the gender gaps and a country's economic and technical development [54, 52], we find that the link is not trivial. The relationship between countries' Internet penetration and the gender gap in OLLE is *not monotonic*, and only when there is a sufficiently high level of women participation in civic society does the OLLE gender gap align with countries' Internet access. In countries with a low level of women's civic participation, the OLLE gap favoring men persists even with the rise of the overall Internet penetration (Fig. 4B). This finding highlights the crucial social condition that allows more literate women to participate online. We also observe non-trivial relationships among multiple inequalities in our analysis of within-country regional disparity. The regional disparity in OLLE is positively associated with unequal education, but the relationship is not simple when comparing countries with different levels of income inequality – for example, a lower disparity OLLE may reflect the homogenized language skills from only the economically advantaged subpopulations, or those from only more educated subpopulations. Our study explicates the complex relationship between the multidimensional inequality measurements and their manifestation on digital populations' online literacy skills.

5.1 Limitations and future research opportunities

We discuss the limitations of this study and highlight where study results must be interpreted with caution, as well as future research opportunities.

Self-selection bias in Facebook data Traditional literacy surveys are expensive to implement and many areas of the world have limited resources for survey research. The challenge for gathering nationally representative samples is not unique to traditional survey research; more recent assessments – for example, PIAAC, which was predominantly administered on computers, were subject to selection effects and therefore required additional adjustment [62]. Our analyses are not immune from self-selection bias where the use of Facebook varies in popularity across different demographics and the differences also vary with countries and regions [63], as well as user subcultures. For example, the observed gender gaps may be due to the over-representation of more privileged women online [64, 65]. On Facebook, a user's comfort level of posting likely depends on their language skills; those with very limited vocabulary may not be in the data, or may choose to communicate via other modalities, e.g., images or videos. This study only considers users' text-based interactions on Facebook and thus the estimates likely miss out on people at the low end of vocabulary skills. To some extent, sampling bias may be mitigated by post-sampling weightings with demographic information, as has been demonstrated in recent data-driven studies [66]. Such an approach nevertheless depends on sufficiently rich demographic information in the data. In our study, only data

disaggregated by gender and coarse-grained geographical grouping are available. Future work may consider tackling the selection bias by separately collecting users' information on demographics and their social media interaction practice.

Representative languages and language-based calibration In the current study, a country's online language literacy was measured based on a single representative language (either the official language or the most used language). One potential risk of relying on a single representative language is that the regional disparity measure in a multilingual country may simply capture the distribution of languages, rather than the diversity of language skills. To address this concern, we perform robust checks in the *Supplemental Material*: S1.5 and Fig. S8 and do not see systematic biases associated with different penetration rates of the representative language.

Another potential risk is to underestimate the language literacy of countries that have sizable language minorities (including people who use/speak a dialect), multilingual communities, or multiple monolingual subpopulations, since their data are largely excluded from our methodology. For example, an English-majority country with a larger Spanish-speaking population may score lower in a measure of English-language literacy skills. For such countries, focusing on improving a dominant-language literacy measure can be potentially harmful, since more resources may be allocated in favor of the dominant language. In *Supplemental Material*: S1.4, we present a case study using India as an example of a multilingual country and show that literacy estimation based on multiple languages has neglectable improvement over English-based estimation in its correlation with the official literacy data (see Fig. S8). However, we acknowledge the official literacy data often have a bias against language minorities and recommend future work consider measuring the online language skills separately for all languages used by sizable populations within a country, to better understand the literacy skills and needs across diverse communities.

The use of official literacy data for cross-language `OLLE` calibration also introduces potential biases and noises. As mentioned in Section 4.1, the post-calibration `OLLE`'s for the Russian-speaking countries are likely to be overestimated due to their historically high literacy rates in the official data. On the other hand, `OLLE`'s for Arabic-speaking countries be underestimated due to the fact that the alternative learning (e.g., religious education) provided in those countries was not included in the official literacy data. Languages concentrated in only one or a few countries, such as Japanese and Korean, are not considered in our study due to the lack of benchmark data that can be used for validation or calibration. Therefore, to establish an adequate common scale for more languages, future research will benefit from more comprehensive and up-to-date data for literacy skills across languages.

Thresholding vs. continuum measurement Our measure relies on thresholding the observed word frequency bands – i.e., the set of LoFF words was identified by the automatically determined word frequency cut-offs – but one may also consider the continuum of the word frequency range. Our choice of focusing on particular word frequency bands is aligned with the existing literature in language comprehension research. For example, studies from English language comprehension distinguish the utility of high-, mid-, and low-frequency vocabulary: the high-frequency vocabulary (e.g., the most frequent 2000- or 3000-word families from a particular English

corpus) provides the largest lexical coverage of any text but is not sufficient for adequate reading comprehension, while the low-frequency vocabulary (including the words over the 9000-word families) is too infrequent and thus has very limited utility; only the mid-frequency vocabulary gives the important range of words required for reading authentic materials [45, 67]. However, different vocabulary sets may serve significant functions for different populations; for example, high-frequency vocabulary has been shown as an important source of knowledge for second-language learners [67]. Future work may take into account the continuum of the word frequency range and investigate the level of contribution provided by the various word frequency bands to online language skills.

Heterogeneity in social media texts A potential concern about using social media text to measure the language skills of a population is how to deal with social media users' heterogeneous behaviors, e.g., some users may post more than others, and some tend to copy content from elsewhere, which could disproportionally impact the population-level measurement. We adopt a few methods to address this concern, including counting each unique unigram once per user, and leaving out posts that are likely to be copy-pasted (see *Supplemental Material*: S1.3 for more details). However, we did not perform efficacy evaluation for these methods, and would encourage future work further examine the impact of text recycling and text production disparities for online literacy assessment.

Aggregate vs. individual measures, and correlations `OLLE`'s are generated based on aggregate data, which inherently poses risks of ecological fallacy compared to other literacy data collected through individual-level tests and surveys. In our study, the between-country correlations only involve the between-country differences in aggregate statistics of the within-country distributions, and the unmeasured within-country measures could be uncorrelated, or could even be correlated in the opposite direction. Taking into account individual assessment in a multi-level analysis with a proper privacy protection mechanism may be a fruitful direction to reduce aggregation bias and the ecological fallacy in future research. In the case of the observed association between the gender gap in `OLLE` and women's civic engagement, a less ambiguous interpretation – whether higher literacy empowers women for civic engagement, or civic engagement leads to legal and institutional changes that enhance literacy, or other cultural, religious, political, and socio-economic conditions influence both women's civic engagement and progress in online language literacy – requires further research to carefully examine the causal pathways.

5.2 Conclusions

This work develops a scalable language literacy measurement to monitor the collective language literacy of the online population using social media data from more than 160 countries. The measure then allows for tracking the trends and inequalities in online language literacy and their relationships with various socioeconomic conditions. Our findings identify key regions and populations disproportionally impacted by literacy challenges, and suggest that education or technical infrastructure alone is not sufficient to explain the variance in online population language literacy

skills. Our study calls out the need for more attention and resources to be allocated to populations with limited online literacy skills – especially those who also suffer from poverty, low resource, and other structural discrimination, to empower them through global challenges such as misinformation and social inequality, and to sustain the overall progress in democratic and socioeconomic development.

Abbreviations
OLLE: online language literacy estimate
LoFF words: lower-frequency words
UNESCO: United Nations Educational, Scientific and Cultural Organization
NALS: National Adult Literacy Survey
PIAAC: Program for International Assessment of Adult Competencies
ELLs: English Language Learners
VST: Vocabulary Size Test
CDF: cumulative distribution function
OLS: ordinary least squares
GII: Gender Inequality Index

Availability of data and materials
Data aggregated at the country level (country-level literacy estimates and summary statistics) will be made available in the Open Science Framework (OSF, at `https://osf.io/zcpej/`) upon publication of this manuscript. Facebook requires that this work was to be done in compliance with Facebook's Data Policy and research ethics review process (`www.facebook.com/policy.php`). Restrictions apply to the availability of the disaggregated data (user- or post-level data), so they are not publicly available. Data aggregated at the country level and other datasets that support the findings of this study will be available from the OSF repository with the permission of the authors, upon reasonable request. The analysis code used to derive the main results are available in the Open Science Framework (OSF) upon publication of this manuscript.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
YRL and SW conceived and designed the research; YRL and SW developed the analysis tools; YRL performed the experiments and analyzed the data; YRL, SW, and WM wrote the paper.

Acknowledgements
We thank anonymous reviewers for their valuable feedback.

Author details
[1]School of Computing and Information, University of Pittsburgh, Pittsburgh, PA, USA. [2]AImpower.org, Mountain View, CA, USA. [3]Meta, Menlo Park, CA, USA.

References
1. Kirsch, I.S., Jungeblut, A., Jenkins, L., Kolstad, A.: Adult literacy in america: A first look at the findings of the national adult literacy survey (2002). NATIONAL CENTER FOR EDUCATION STATISTICS. Access on 09/23/2022 at `https://nces.ed.gov/pubs93/93275.pdf`
2. Kutner, M., Greenberg, E., Jin, Y., Boyle, B., Hsu, Y.-c., Dunleavy, E.: Literacy in Everyday Life: Results from the 2003 National Assessment of Adult Literacy. NCES 2007-490. Report, U.S. Department of Education. Washington, DC: National Center for Education Statistics (2007)
3. Schütz, G., Ursprung, H.W., Wößmann, L.: Education policy and equality of opportunity. Kyklos **61**(2), 279–308 (2008)
4. Kutner, M., Greenberg, E., Jin, Y., Boyle, B., Hsu, Y.-c., Dunleavy, E., White, S.: Literacy in everyday life: Results from the 2003 national assessment of adult literacy (2017). (Accessed on 09/23/2022)
5. Ferrer, A., Green, D.A., Riddell, W.C.: The effect of literacy on immigrant earnings. Journal of Human Resources **41**(2), 380–410 (2006)
6. Bonikowska, A., Green, D.A., Riddell, W.C.: Literacy and the Labour Market: Cognitive Skills and Immigrant Earnings. Statistics Canada, Ottawa (2008)
7. Schwerdt, G., Wiederhold, S., Murray, T.S.: Literacy and growth: New evidence from PIAAC. Retrieved from PIAAC Gateway website: `http://piaacgateway.com/`. (Accessed on 06/30/2020) (2020)
8. Dewalt, D., Berkman, N., Sheridan, S., Lohr, K., Pignone, M.: Literacy and health outcomes: A systematic review of the literature. Journal of general internal medicine **19**, 1228–39 (2005). doi:10.1111/j.1525–1497.2004.40153.x
9. Desjardins, R., Thorn, W., Schleicher, A., Quintini, G., Pellizzari, M., Kis, V., Chung, J.E.: OECD Skills Outlook 2013: First Results from the Survey of Adult Skills. `http://dx.doi.org/10.1787/9789264204256-en`. Accessed on 11/23/2020 (2013)
10. Gerger, C.: Social linguistics and literacies: Ideology in discourses. social linguistics and literacies: Ideology in discourses. Ilha do Desterro (2008)
11. UNESCO Institute for Statistics: Literacy Rates Continue to Rise from One Generation to the Next. Retrieved September 22, 2022 from `http://uis.unesco.org/sites/default/files/documents/fs45-literacy-rates-continue-rise-generation-to-next-en-2017_0.pdf` (2017)

12. Mundial, G.B., UNICEF, et al.: Education 2030: Incheon declaration and framework for action: towards inclusive and equitable quality education and lifelong learning for all (2016)
13. Bach, A.J., Wolfson, T., Crowell, J.K.: Poverty, literacy, and social transformation: An interdisciplinary exploration of the digital divide. Journal of Media Literacy Education **10**(1), 22–41 (2018)
14. Relations, M.I.: Meta Reports Third Quarter 2022 Results. https://investor.fb.com/investor-news/press-release-details/2022/Meta-Reports-Third-Quarter-2022-Results/default.aspx (2022)
15. Relations, M.I.: Meta Earnings Presentation Q3 2022. https://s21.q4cdn.com/399680738/files/doc_financials/2022/q3/Q3-2022_Earnings-Presentation.pdf (2022)
16. Rammstedt, B., Maehler, D.B.: Introduction: PIAAC and its Methodological Challenges. methods, data, analyses **8**(2), 12 (2016)
17. SDG17: United Nations Sustainable Development Goals. https://sdgs.un.org/goals. Accessed on 10/15/2020 (2015)
18. Montoya, S.: 50 Years of International Literacy Day: Time to Develop New Literacy Data — UNESCO UIS. http://uis.unesco.org/en/blog/50-years-international-literacy-day-time-develop-new-literacy-data. (Accessed on 06/30/2020) (2016)
19. UNESCO Institute for Statistics: UIS Statistics. Retrieved September 22, 2022 from http://data.uis.unesco.org/index.aspx?queryid=3445# (2022)
20. NCES: Program for the International Assessment of Adult Competencies (PIAAC). https://nces.ed.gov/surveys/piaac/. (Accessed on 09/23/2022) (2012)
21. Hargittai, E.: An update on survey measures of web-oriented digital literacy. Social science computer review **27**(1), 130–137 (2009)
22. DiMaggio, P., Hargittai, E., *et al.*: From the "digital divide" to "digital inequality": Studying internet use as penetration increases. Princeton: Center for Arts and Cultural Policy Studies, Woodrow Wilson School, Princeton University **4**(1), 4–2 (2001)
23. Sprague, K., Grijpink, F., Manyika, J., Moodley, L., Chappuis, B., Pattabiraman, K., Bughin, J.: Offline and falling behind: Barriers to Internet adoption. Retrieved September 22, 2022 from https://www.mckinsey.com/~/media/mckinsey/dotcom/client_service/high%20tech/pdfs/offline_and_falling_behind_full_report.ashx (2014)
24. Ortiz-Ospina, E.: The rise of social media - Our World in Data. https://ourworldindata.org/rise-of-social-media. (Accessed on 11/20/2021) (2019)
25. Erstad, O., Gilje, N., de Lange, T.: Re-mixing multimodal resources: Multiliteracies and digital production in norwegian media education. Learning, Media and Technology **32**, 183–198 (2007). doi:10.1080/17439880701343394
26. Alvermann, D.E.: Why bother theorizing adolescents' online literacies for classroom practice and research? Journal of Adolescent & Adult Literacy **52**(1), 8–19 (2008). Accessed 2022-10-01
27. Greenhow, C., Robelia, B.: Old communication, new literacies: Social network sites as social learning resources. J. Computer-Mediated Communication **14**, 1130–1161 (2009). doi:10.1111/j.1083-6101.2009.01484.x
28. Perkel, D.: Copy and paste literacy? literacy practices in the production of a myspace profile. Informal Learning and Digital Media **49** (2010). doi:10.1590/S0103-18132010000200011
29. Greenhow, C., Gleason, B.: Twitteracy: Tweeting as a new literacy practice. The Educational Forum **76**(4), 464–478 (2012). doi:10.1080/00131725.2012.709032
30. Davies, J.: Facework on facebook as a new literacy practice. Computers & Education **59**(1), 19–29 (2012). doi:10.1016/j.compedu.2011.11.007. CAL 2011
31. Kress, G.: Literacy in the new media age, 1–190 (2003). doi:10.4324/9780203299234
32. Clark, C., Dugdale, G.: People's Writing: Attitudes, behaviour and the role of technology. https://files.eric.ed.gov/fulltext/ED510271.pdf. (Accessed on 09/30/2022) (2009)
33. Sabaruddin: Facebook utilisation to enhance english writing skill. English Language Teaching **12**(8), 37–43 (2019)
34. Black, R.W.: Just don't call them cartoons: The new literacy spaces of anime, manga, and fanfiction. In: Coiro, J., Knobel, M., Lankshear, C., Leu, D.J. (eds.) Handbook of Research on New Literacies, pp. 583–610. Taylor & Francis, New York (2008)
35. Kojo, D.B., Agyekum, B.O., Arthur, B.: Exploring the effects of social media on the reading culture of students in tamale technical university. Journal of Education and Practice **9**(7), 47–56 (2018)
36. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018). doi:10.1126/science.aap9559. https://www.science.org/doi/pdf/10.1126/science.aap9559
37. Edelson, L., Nguyen, M.-K., Goldstein, I., Goga, O., McCoy, D., Lauinger, T.: Understanding engagement with u.s. (mis)information news sources on facebook. In: Proceedings of the 21st ACM Internet Measurement Conference. IMC '21, pp. 444–463. Association for Computing Machinery, New York, NY, USA (2021). doi:10.1145/3487552.3487859. https://doi.org/10.1145/3487552.3487859
38. OECD: 21st-Century Readers: Developing Literacy Skills in a Digital World. OECD Publishing, https://www.oecd-ilibrary.org/content/publication/a83d84cb-en (2021)
39. Lee, J.: Size matters: Early vocabulary as a predictor of language and literacy competence. Applied Psycholinguistics **32**(1), 69 (2011)
40. Curtis, M.E.: The role of vocabulary instruction in adult basic education. Comings, J., Garner, B., Smith, C., Review of Adult Learning and Literacy **6**, 43–69 (2006)
41. Ouellette, G.P.: What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. Journal of educational psychology **98**(3), 554 (2006)
42. National Research Council: Improving Adult Literacy Instruction: Options for Practice and Research. National Academies Press, Washington, DC (2012). doi:10.17226/13242
43. Schmitt, N.: An Introduction to Applied Linguistics. Routledge, New York (2013)
44. Laurén, C., Nordman, M.: Special Language: From Humans Thinking to Thinking Machines. Multilingual

Matters, Clevedon, Philadelphia (1989)
45. Nation, I.: How large a vocabulary is needed for reading and listening? Canadian modern language review **63**(1), 59–82 (2006)
46. Beglar, D., Nation, P.: A vocabulary size test. The language teacher **31**(7), 9–13 (2007)
47. Grave, É., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning word vectors for 157 languages. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)
48. Goldberg, Y., Orwant, J.: A dataset of syntactic-ngrams over time from a very large corpus of english books. In: Second Joint Conference on Lexical and Computational Semantics, Atlanta, Georgia, USA, pp. 241–247 (2013)
49. Satopaa, V., Albrecht, J., Irwin, D., Raghavan, B.: Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In: 2011 31st ICDCSW, pp. 166–171 (2011). IEEE
50. Antunes, M., Gomes, D., Aguiar, R.L.: Knee/elbow estimation based on first derivative threshold. In: 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService), pp. 237–240 (2018). IEEE
51. Beasley, T.M., Erickson, S., Allison, D.B.: Rank-based inverse normal transformations are increasingly used, but are they merited? Behavior genetics **39**(5), 580 (2009)
52. Fatehkia, M., Kashyap, R., Weber, I.: Using Facebook ad data to track the global digital gender gap. World Development **107**, 189–209 (2018)
53. Roser, M., Ritchie, H., Ortiz-Ospina, E.: Internet. Our World in Data (2022). https://ourworldindata.org/internet
54. World Economic Forum: The global gender gap report. (2020). World Economic Forum Genebra
55. OECD: PISA 2018 Results (Volume II), p. 376 (2019). doi:10.1787/b5fd1b8f-en. https://www.oecd-ilibrary.org/content/publication/b5fd1b8f-en
56. United Nations Development Programme: Technical Notes: Calculating the human development indices—graphical presentation. Retrieved from UNDP website: `https://hdr.undp.org/sites/default/files/2021-22_HDR/hdr2021-22_technical_notes.pdf`. (Accessed on 10/26/2022) (2021)
57. Coppedge, M., Gerring, J., Knutsen, C.H., Krusell, J., Medzihorsky, J., Pernes, J., Skaaning, S.-E., Stepanova, N., Teorell, J., Tzelgov, E., *et al.*: The methodology of "varieties of democracy"(v-dem). Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique **143**(1), 107–133 (2019)
58. Warf, B.: Geographies of the Internet. Routledge, New York (2020)
59. World Bank: World development report 2016: Digital dividends. (2016). Washington, DC: World Bank
60. Zickuhr, K., Smith, A.: Digital differences. Retrieved December 7, 2020 from `https://www.pewresearch.org/internet/2012/04/13/digital-differences/` (2012)
61. Hilbert, M.: Digital gender divide or technologically empowered women in developing countries? a typical case of lies, damned lies, and statistics. Women's Studies International Forum **34**(6), 479–489 (2011). doi:10.1016/j.wsif.2011.07.001
62. Yamamoto, K., Khorramdel, L., Von Davier, M., et al.: Scaling PIAAC Cognitive Data
63. Facebook: Facebook Q2 2020 Earnings. `https://s21.q4cdn.com/399680738/files/doc_financials/2020/q2/Q2-2020-FB-Earnings-Presentation.pdf`. (Accessed on 10/15/2020) (2020)
64. Magno, G., Weber, I.: International gender differences and gaps in online social networks. In: International Conference on Social Informatics, pp. 121–138 (2014). Springer
65. Kashyap, R., Verkroost, F.C.: Analysing global professional gender gaps using linkedin advertising data. EPJ Data Science **10**(1), 39 (2021)
66. Park, M., Thom, J., Mennicken, S., Cramer, H., Macy, M.: Global music streaming data reveal diurnal and seasonal patterns of affective preference. Nature human behaviour **3**(3), 230–236 (2019)
67. Masrai, A.: Vocabulary and reading comprehension revisited: Evidence for high-, mid-, and low-frequency vocabulary knowledge. Sage Open **9**(2), 2158244019845182 (2019)

Figures
Figure 1-4.

Additional Files
Additional file — Supplemental Information
This document includes the additional information: (1) Supplementary text; (2) Figures S1 to S7; (3) Tables S1 to S12; (4) SI References. In addition, data aggregated at the country level (country-level literacy estimates and summary statistics) and the R analysis code for deriving the main results will be made available in the Open Science Framework (OSF) upon publication of this manuscript. Facebook requires that this work was to be done in compliance with Facebook's Data Policy and research ethics review process (`www.facebook.com/policy.php`). Restrictions apply to the availability of the disaggregate data (user- or post-level data), so they are not publicly available. Data aggregated at the country level and other datasets that support the findings of this study will be available from the OSF repository with the permission of the authors, upon reasonable request.

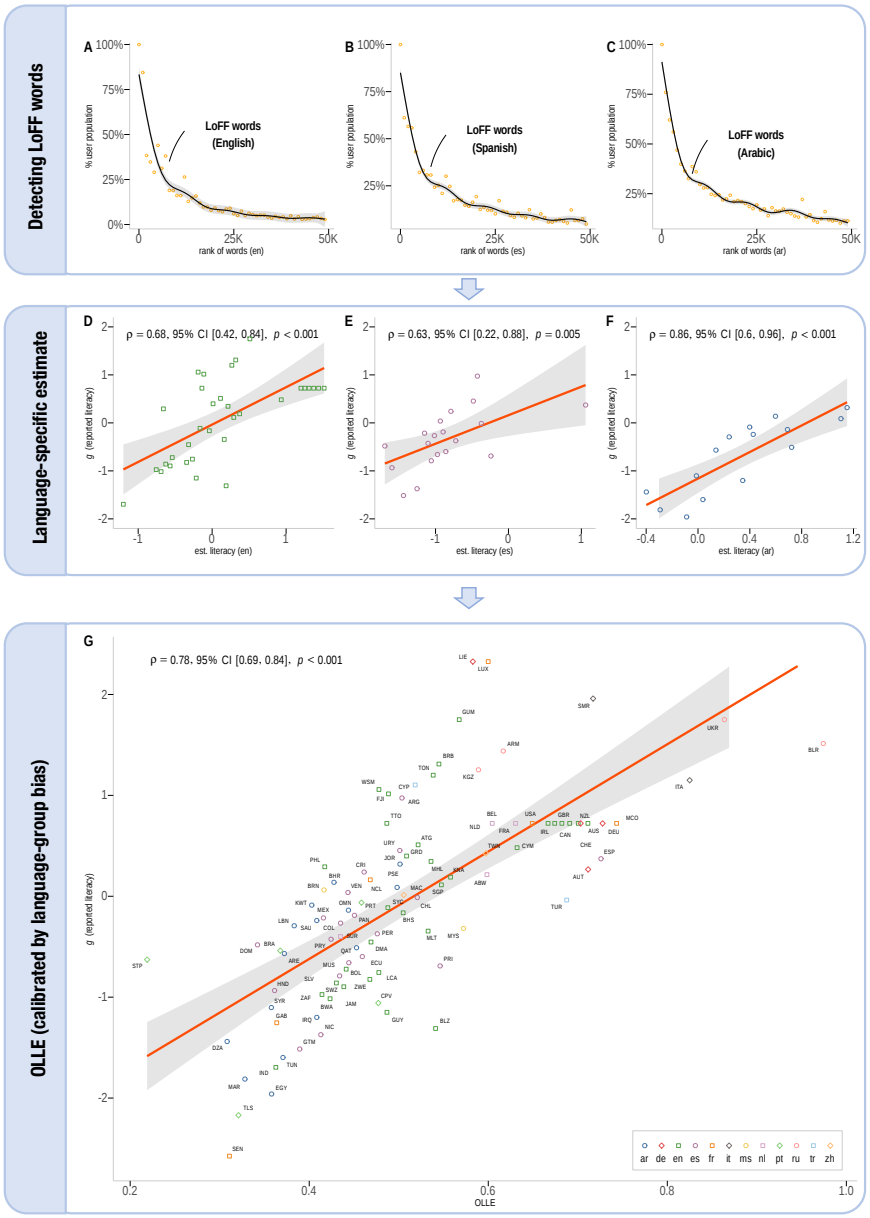

Figure 1: **Creating online language literacy estimate.** Our methodology produces online language literacy estimate (OLLE) through three major steps: 1. (A-C) The set of LoFF words (lower-frequency frequent words) is algorithmically determined based on the vocabulary popularity in the language corpus; the red bands in (A-C) indicate the selected sets of LoFF words in the three most widely used languages in our data (English, Spanish, Arabic). 2. (D-F) Normalized total occurrence of LoFF words in Facebook dataset from each country is used as a language-specific online literacy estimate for that country. (D-F) show the strong correlations found between our estimates and countries' officially reported literacy rates in English, Spanish, and Arabic, respectively. 3. (G) The calibrated global estimates, OLLE, are generated after addressing language group bias and shown here with a strong correlation with reported literacy rates (Spearman's rank correlation coefficient $\rho = 0.78$, 95% CI [0.69, 0.84], $p < 0.001$.) Error bounds represent the 95% confidence intervals. In (D-G), each dot represents a country, with x value indicating the country's raw (D-F) or calibrated (G) literacy estimate and y value the country's officially reported literacy rate.

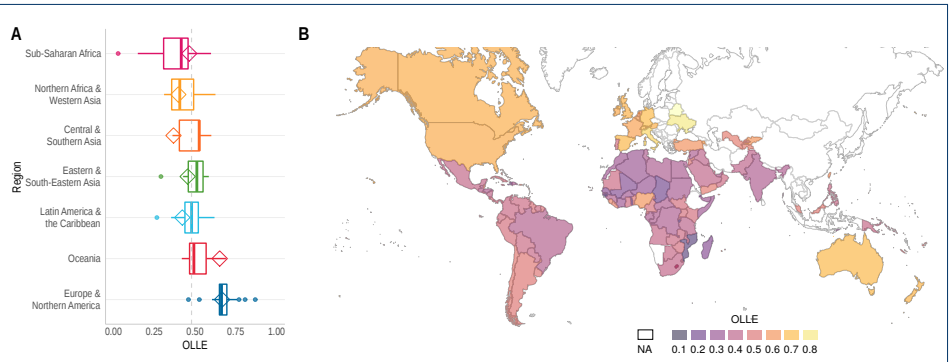

Figure 2: **Online language literacy estimates across the world.** (A) Summary of the `OLLE`'s across the seven geographical groups. The dashed line marks the global average (0.477, population-weighted). Within each box, a line denotes the median value of the group, and a diamond indicates the population-weighted mean of the group; boxes extend from the first to the third quartile of each group's distribution of values; whiskers (lines extending from the boxes) denote the most extreme values within 1.5 interquartile range of each group; dots denote observations outside the range of extreme values. (B) Map of the `OLLE`'s available in our dataset.

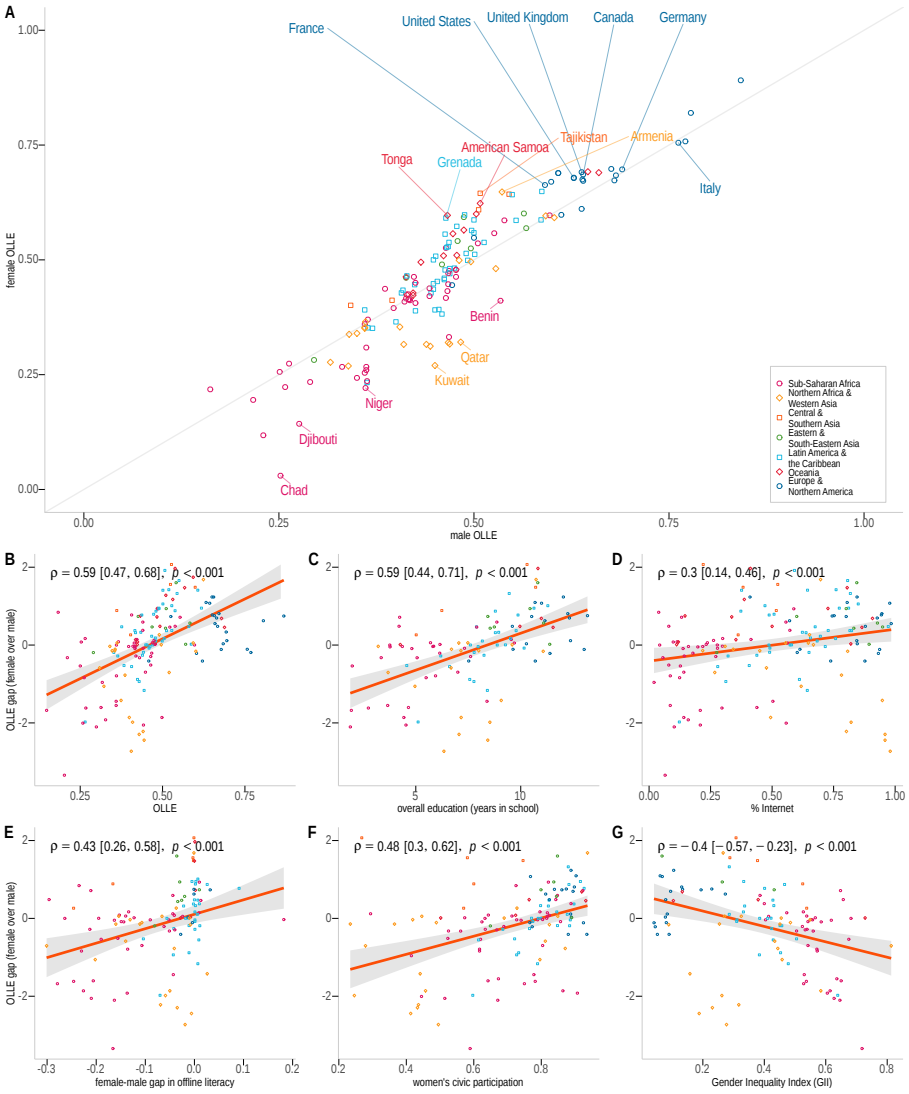

Figure 3: **Gender gap in online language literacy.** (A) Country-level `OLLE`'s for women and men. (B-G) The standardized online literacy gender difference (women over men) compared with the country's overall `OLLE`, education, Internet penetration, as well as gender parity and empowerment measures. Error bounds represent the 95% confidence intervals. Spearman's rank correlation coefficients and 95% CI are indicated in the scatterplots.

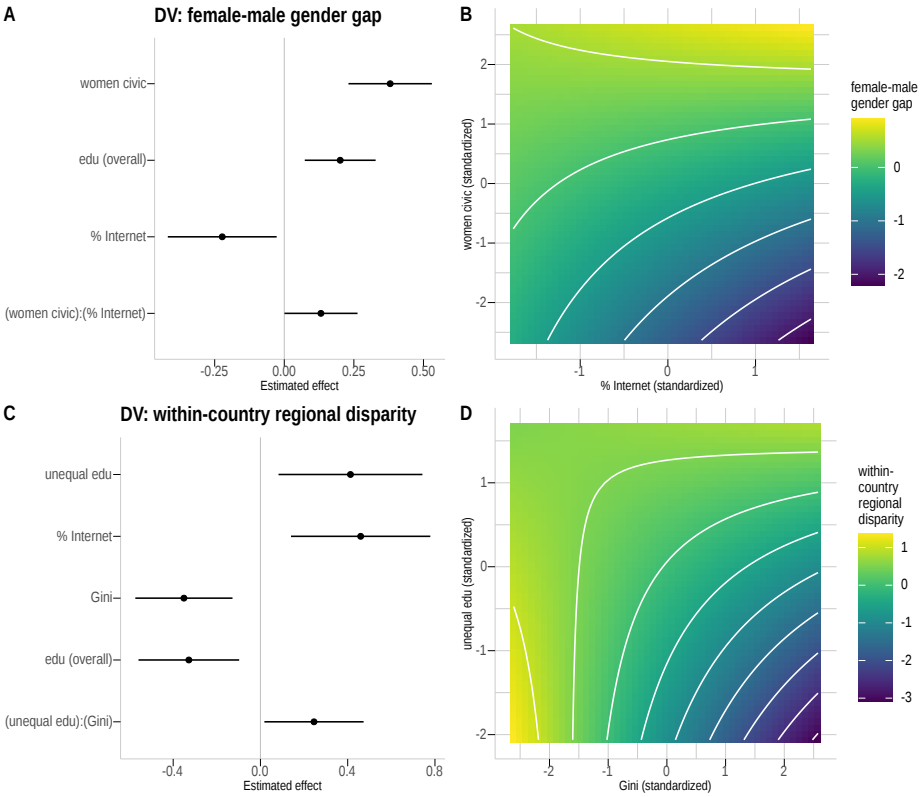

Figure 4: **The societal context and disparities in OLLE.** (A) Multiple regression model shows robust associations of the gender gap (female over male) in OLLE with overall education status and women's civic participation. (B) The significant interaction between women's civic participation and Internet penetration in predicting gender gap in OLLE. (C) Multiple regression models show robust associations of regional disparity in OLLE with inequality in education, internet penetration, inequality in income (Gini index), and overall educational attainment. (D) Inequalities in education and income appear to have the opposite contribution in predicting a country's regional disparity in OLLE. Full regression estimates are provided in *Supplemental Material* Tables S9 and S12, respectively. (B,D) show the average marginal effects of the two interaction terms.