SCIENTIFIC Publishing Limited

Digital Humanities and Society Studies
https://ojs.ukscip.com/index.php/dhss

Article

Epistemological Asymmetries in the Digital Archive: Cultural Bias in LLaMA-1 and CommonCrawl-Based Large Language Models and Its Implications for Digital Humanities Practice

Mayuree Pal [1,*] **and C.P. Rashmi** [2]

[1] School of Media Studies, Presidency University, Bengaluru 560119, India
[2] Amity School of Communication, Amity University, Bengaluru 562110, India
[*] Correspondence: mayureepal18@gmail.com

Received: 11 February 2026; **Revised:** 1 May 2026; **Accepted:** 13 May 2026; **Published:** 25 June 2026

Abstract: Given the increasing integration of large language models (LLMs) in the digital humanities (DH) for textual analysis, archival research, and computational interpretation, an urgent need to decolonize LLM knowledge structures has emerged. This study conducts a secondary data analysis to highlight profound epistemological asymmetries embedded within LLaMA-1 and other documented CommonCrawl-based large language models and digital archives. The quantitative analysis reveals a stark representational disparity: approximately 82% of LLaMA-1's training set is explicitly English-filtered, and 49–56% of top-website content is in English, despite only 25.9% of global internet users speaking English as a first language. After more than a decade of systematic funding and infrastructure development, the digitisation of cultural heritage within European institutions remains stagnant at merely 16–23%. Crucially, no equivalent systematic surveys of cultural heritage digitisation exist for institutions in the Global South, indicating a severe global structural deficit. To mitigate these cascading disparities (from biased training data to uneven archival digitisation), this paper proposes the Epistemological Asymmetry Framework (EAF). The EAF serves as a structured, verifiable tool designed for DH practitioners and research institutes to audit, measure, and actively address cultural and linguistic biases. Ultimately, establishing equitable digital practices requires both rigorous multilingual corpus auditing and significant infrastructural investments in global heritage digitisation.

Keywords: Large Language Models; Cultural Bias; Digital Heritage; ENUMERATE; Multilingual Corpora; LLaMA; W3Techs

1. Introduction

Representation has always been central to digital humanities: whose texts are kept, whose voices are collected, whose cultural production documents the record. These questions were further amplified, not resolved, by the computational turn. As large language models (LLMs) like ChatGPT increasingly leverage archives for tasks such as close-reading, automated content analysis, cross-language translation, and understanding and interpreting collections remotely stored, the representational politics of the archive has now become embodied in the weights and training distributions of the machine learning systems.

The focus of this paper is on one concrete and empirically accessible problem, which is the lack of fit between the training sets of LLMs and the actual diversity of human heritage that the field of Digital Humanities seeks to serve

as a beacon for. In contrast to most questions about opacity in AI, the critical information necessary for describing this question is public and can be quoted directly. The complete breakdown of the training corpus composition is given in the LLaMA-1 paper. Content language information in the top 10 million websites is reported and updated by W3Techs [1]. The Europeana ENUMERATE project has published four consecutive reports on the current state of digitisation in the cultural heritage sector, with a common result of insufficient progress.

This paper aims to bring these three well-recognized datasets to harmoniously support each other, and to demonstrate the extent and configuration of the epistemological asymmetry revealed through the dataset and to propose a feasible model, Epistemological Asymmetry Framework (EAF), that can be used by scholars and organizations to audit and actively engage in solving these asymmetries. All statements in the paper are clearly supported by the primary data cited. When qualitative interpretation is provided, it is explicitly different from the quantitative evidence. The study addresses three explicit research questions: (RQ1) To what extent does the training corpus of LLaMA-1 over-represent English-language content relative to its share of global web content and internet user population? (RQ2) What is the baseline rate of cultural heritage digitisation in Europe's best-resourced sector, and what does the absence of equivalent data for the Global South indicate? (RQ3) How can a structured, verifiable framework (EAF) be constructed to help DH practitioners audit and address epistemological asymmetries in LLM-mediated archival work?

2. Conceptual Background

2.1. Training Data and Epistemic Exclusion

There is a well-known correlation between the structure of training data and model behaviour. Moreover, LLMs are trained on data and learn what they are exposed to during training due to statistical patterns in the data [2,3]. It is not a subtle or intensely debatable point, it is the real soil that one focuses on and writes down in choosing and in documenting one's corpus. When Touvron et al. (2023) publish **Table 1** of the LLaMA paper showing that 67% of training data consists of English-filtered CommonCrawl and a further 15% is the English-only C4 corpus, they are documenting an architectural decision with direct epistemic consequences: the model's competencies, associations, and cultural framings will be shaped overwhelmingly by English-language, US-centric web content [4–6].

Table 1. Content Language Distribution of Top 10 Million Websites (W3Techs, 2026) with Internet User Context.

Language	Share of Top-10 Million Websites (May 2023)	Share of Top-10 Million Websites (Dec. 2025)	Contextual Note
English	55.5%	49.3%	Still the single largest language bloc despite decline; LLaMA-1 training corpus is 82% English (CommonCrawl 67% + C4 15%) [4]
Spanish	5.0%	6.0%	Growing presence; moderately covered in multilingual LLMs; C4 corpus is English-only
German	4.3%	5.9%	Well-represented in European language models (e.g., BLOOM, multilingual BERT)
Japanese	3.7%	5.1%	Japan has independently developed Japanese-specific LLMs (e.g., Rinna) in response to cultural and linguistic misrepresentation
French	4.4%	4.5%	Relatively stable; LLaMA-1 explicitly supports 20 European languages including French
Russian	4.9%	3.7%	Declining share; Cyrillic script broadly supported in LLaMA-1's 20-language European set
Chinese	1.4%	1.1%	Internet users: 19.4% of global total (Internet World Stats/Statista, 2020) [7]. Gap between user share and content share is extreme
Arabic	0.7%	0.5%	Internet users: 5.2% of global total (Internet World Stats/Statista, 2020) [7]. Declining content share despite growth in users
Indonesian/Malay	0.7%	1.0%	Internet users: 4.3% of global total (Internet World Stats/Statista, 2020) [7]. Slight content growth but user-content gap persists
Hindi/South Asian scripts	<0.1%	<0.1%	Devanagari script represents 0.1% of all web scripts [1]. Hindi internet users: estimated 14.5% of global users by OBDILCI [8]. Note: W3Techs reports script share in web content; OBDILCI reports internet user share—these measure different phenomena. Extreme gap.
All remaining languages	<0.1% each	<0.1% each	Covers 7,000+ world languages including all Sub-Saharan African, Indigenous, Pacific, and most South/Southeast Asian languages

Source: Content language data from 'Historical trends in the usage statistics of content languages for websites', W3Techs (December 2025, cited 2026) [1]—measures language of web content only. Internet user population percentages (e.g., 19.4% Chinese, 5.2% Arabic) from Internet World Stats/Statista (2020) [7]—measures language of internet users, a distinct dataset. Hindi user estimate from OBDILCI [8]. LLaMA training data from Touvron et al. (2023) [4]. These sources measure different phenomena and should not be directly conflated.

According to an analysis released by Baack (2024), out of the 47 text-generating LLMs that have been published

since 2019, at least 64% were created using data from Common Crawl that had some filtering applied, and that such builders 'essentially declare a relatively small subset of mostly English web pages as representative of the whole world' [9]. The C4 corpus, which is 750GB worth of English text, is only explicitly used during the training of LLaMA-1, T5, and many more models, according to the documentation. All of them are not inferences; rather, they are documented design decisions [10].

2.2. Web-Based Language Distribution

The continuity of content language statistics released by W3Techs since 2011 and cited in academic papers can be used as a good touchpoint to measure LLM training data. By May 2023, 55.5% of the 10 million most popular websites were in English; this dropped to 49.3% by December 2025 [1]. These numbers show the measurable shift of dominance towards English, which relates to the deliberate curation, with the training dataset of LLaMA-1 having ~82% English dominance as compared to the top-website content of ~49–56% [11–13].

The W3Techs data also indicate the position of languages spoken by less digitally resourced communities. In addition, the Chinese language content included in the Top websites is relatively small (1.1%–1.4%) despite there being 19.4% Chinese internet users out of approximately 4.66 billion global internet users recorded in 2020 [7]. It is important to note that W3Techs measures web content language distribution, while internet user language data is drawn from Internet World Stats and Statista (2020) [7] —these are distinct datasets measuring different phenomena. The estimated 14.5% of internet users who speak Hindi [8], in comparison, are currently only served by less than 0.1% of available website content, and Hindi data written in Devanagari script is similarly scarce from currently available standard web corpora. As per W3Techs, Arabic is the native language of approximately 420 million people and the language is used only 0.5% of the website content [1].

2.3. Digital Heritage Digitisation and the European Baseline

ENUMERATE is a series of Core Surveys carried out by the Europeana Foundation that provides the most systematic longitudinal information on the state of the art of digitisation of cultural heritage institutions. This series has traced approximately 1,000–1,400 institutions in Europe and reveals that ten years on, only 20% (in 2012), 17% (in 2014), 23% (in 2015) and 16% (in 2017) of collections had undergone the digitisation process as a result of EU investment in digitisation [14]. On the world stage these are the best of times, with a financial baseline, EU policy support, and technical infrastructure in place. This is significant enough in its own right because there is no baseline survey available for institutions in sub-Saharan Africa, South Asia, the Middle East or Latin America comparable to this one, suggesting the global baseline for the Global South might be much lower. The lack of systematic web and archival preservation inherently biases historical research [15].

In 2009, the UN Education, Science and Culture Organisation (UNESCO) developed and accepted its Charter on the Preservation of Digital Heritage, stating that the preservation of digital heritage is a global responsibility [16]. However, unfortunately, taking this aim into practice in regular monitoring outside of Europe has still not been achieved. The latest UNESCO IFAP programme reports (2024) mention impressive pilot-scale digitisation projects like the 5,000+ audio recordings in Libya or the development of national archive systems in Sri Lanka, but these are yet partial digitisation projects and not systematic surveys [17]. This absence of data is itself data.

3. Methodology

3.1. Research Design: Secondary Data Analysis

Secondary data analysis of three primary datasets publicly available was used for this study. Secondary data analysis can be used when existing data has already been published, peer-reviewed, and validated by institutions or processes; and when the research questions pertained to structural patterns within existing documented systems which it would be inappropriate to replicate by primary data collection. The three data sets chosen meet the verifiability requirements: All numbers used in the tables in this paper were extracted from the sources indicated, without any interpolation or estimation.

3.1.1. Dataset 1: W3Techs Content Language Statistics

W3Techs conducts an analysis based on Tranco's top 10 million websites and reports the language used on their main page. The statistics are publicly available and continuously updated. The snapshots used are those of May 2023 (English 55.5%) and December 2025 (English 49.3%) published on the W3Techs site [4]. There are limitations to the W3Techs methodology that are found in the academic literature: The language is only picked up on the home page of a site; a multilingual site is considered a single-language site; only the top 10 million websites are covered. These are referred to, if applicable.

3.1.2. Dataset 2: LLaMA-1 Training Data Composition

The LLaMA-1 paper was published by the authors Touvron et al. (2023) on arXiv, and is fully disclosed regarding the composition of its training data (**Table 1** of the paper) [4]. This table summarizes 7 data sources, sampling rate, number of epochs, and disk size. The criteria for selection are that LLaMA-1 is (a) publicly released with documentation, (b) explicitly releases the composition of the training set in tabular form and (c) represents the LLM training set corpus design space. Baack (2024) reports that 64% of 47 LLMs released between 2019–2023 rely on at least one subset of Common Crawl for training data, specifically the T5 version of CC which is 82% from Common Crawl. LLaMA-1 is not a unique example, rather, an exemplar [9].

3.1.3. Dataset 3: ENUMERATE Core Survey Series

The ENUMERATE Core Survey series is comprised of four surveys directed at European cultural heritage institutions which took place in 2012, 2014, 2015 and 2017. All the survey reports are freely available from the Europeana Foundation website [14]. Specific figures used for this study are given in the main findings of each survey in percentages of institutions that have a digital collection, collections digitised, collections that require digitisation, and the percentage that have a written digital strategy. The objective of using the ENUMERATE data in this research is to establish a maximum likely digitisation rate in the world's best resourced cultural heritage sector and thus provide a context for the global picture.

3.2. Analytical Framework Construction

Epistemological Asymmetry Framework (EAF) is developed inferentially by using the three datasets. One of the dimensions in the framework is an analytical question which can be answered at least in part based on specifically cited data relating to particular data sets. The framework is provided in **Table 2**. The manuscript explicitly frames the LLaMA-1/W3Techs corpus data and the ENUMERATE digitisation data as complementary diagnostic dimensions of a shared structural asymmetry rather than as causally linked phenomena: no claim is made that low digitisation rates cause the English dominance of LLM training corpora, or vice versa; each dataset diagnoses a distinct layer of the same epistemological problem. The purpose is to provide not an exclusive list of topics, but a useful vocabulary to structure the audit efforts of researchers and institutions. If there is no systematic reasonable verifiable data available, for example, estimates of the proportion of digitisation in non-European countries, the information will be recorded as 'data absent', which is also a valid result.

Table 2. Epistemological Asymmetry Framework (EAF): Dimensions, Verified Indicators, and DH Interventions.

EAF Dimension	Primary Data Source	Verifiable Indicator	Recommended DH Intervention
I. Linguistic Coverage in Training Data	Touvron et al. (2023) [4]; W3Techs infrastructure context	Percentage of English-only or English-dominant sources in training corpus. LLaMA-1: 82% explicitly English-filtered (CommonCrawl 67% + C4 15%).	Mandate multilingual corpus audits before DH deployment. Partner with CLARIN, DOBES, and ELDP for low-resource language data. Require training-data disclosure from LLM developers.
II. Heritage Collection Digitisation	ENUMERATE Core Surveys (Europeana, 2012–2017) [14]	Percentage of European collections digitised: 16-23% across all four survey rounds. Global South comparator data absent which is itself an evidential indicator of structural disadvantage.	Support UNESCO IFAP collaborative digitisation programmes. Fund Southern-led digitisation initiatives. Advocate for non-European ENUMERATE-equivalent surveys to fill the global measurement gap.
III. Institutional Readiness and Access	ENUMERATE Core Survey 4 [14]; W3Techs infrastructure context	Percentage of European institutions with a written digital strategy: 42% as of 2017 (up from 34% in 2015). Global institutional readiness data absent.	Develop offline-capable DH tools for low-bandwidth environments. Provide open-access API tiers. Co-design workflows with local institutions. Promote formal digitisation policy adoption as prerequisite for LLM-augmented DH.

4. Findings

4.1. Web Content Language Distribution (W3Techs)

In **Table 1** the content language distribution of the top 10 million websites is shown at two points in time: May 2023 and December 2025, based on the data of W3Techs (2026). Where possible, internet user population figures are provided from Internet World Stats/Statista (2020) [7] to give context to the disparity between internet usage population and the content creators [1].

Table 1 shows there are two very different asymmetry patterns. The first concerns the English proportion of web content (~49–56%) versus the proportion of English in LLaMA-1's training data (~82%). If someone translates an English-language page into another language, the CommonCrawl pipeline makes it a point to reject all such pages, and the notion of C4 being based on English-language content is intentional. What is more important, is the distinction between the languages of the Global South, and the proportions of users and content representation in those languages [18–20]. 19.4% of Internet users, and 1.1% of the top-website content is Chinese. Approximately 14.5% to 15% of users are Hindi speakers while less than 0.1% of the website is in Hindi.

4.2. LLaMA-1 Training Corpus Composition

Table 3 is the same as **Table 1** of Touvron et al. (2023), with added analyses of the linguistic coverage of each data set [4].

Table 3. LLaMA-1 Pre-Training Data Composition.

Data Source	Sampling Proportion (%)	Epochs	Disk Size	Language/Scope
CommonCrawl (English-filtered)	67.0%	1.10	3.3 TB	English only. Non-English pages explicitly removed via fastText language identification
C4 (Colossal Clean Crawled Corpus)	15.0%	1.06	783 GB	English only. C4 is an English-language corpus derived from the April 2019 Common Crawl snapshot
GitHub	4.5%	0.64	328 GB	Mostly English. Code-centric; reflects demographics of open-source software contributors
Wikipedia	4.5%	2.45	83 GB	20 languages (all European: Latin and Cyrillic scripts only). No Arabic, Chinese, Hindi, Swahili, etc.
Books (Gutenberg + Books3)	4.5%	2.23	85 GB	Predominantly English. Reflects Western literary canon; Global South literatures marginalised
ArXiv	2.5%	1.06	92 GB	English-dominant academic publishing. Reflects Global North research institutions
StackExchange	2.0%	1.03	78 GB	Primarily English. Reflects demographics of English-speaking developer communities
TOTAL English or English-dominant sources	~82–97%	—	—	CommonCrawl (67%) + C4 (15%) = 82% explicitly English-filtered. Wikipedia adds 20 European languages only.

Source: All sampling proportions, epoch counts, and disk sizes are directly reproduced from **Table 1** of Touvron et al. (2023), 'LLaMA: Open and Efficient Foundation Language Models' [4]. The contextual 'Language/Scope' column reflects documented properties of each corpus. CommonCrawl filtering described in Touvron et al. (2023) [4]: 'language identification with a fastText linear classifier to remove non-English pages.' C4 language scope from Raffel et al. (2019) [10]. LLaMA Wikipedia scope (20 European languages, Latin and Cyrillic scripts) from Touvron et al. (2023) [4].

However, to know the proportion of such a language in the data that a sample is drawn from, the training data of each LLaMA model has to be examined. The percentage of English or English-dominant sources in the total training data is shown in the figures in **Table 3**; the data is English (Commoncrawl 67%) plus C4 (15%). Within the remaining 18%, the Wikipedia subset covers 20 European languages written exclusively in Latin or Cyrillic scripts (with no Arabic, Chinese, Hindi or Swahili), while the academic repository ArXiv and the developer forum StackExchange are both English-dominant. Considering the full training set, less than 5% of LLaMA-1's data originates from outside European language traditions and their colonial extensions.

LLaMA-1 is not an isolated case: Baack (2024) reports that at least 64% of the 47 text-generating LLMs published between 2019 and 2023 relied on filtered versions of Common Crawl [9]. No comparable training-composition table has been published for closed models such as GPT-4, so the scoped claims of this paper are restricted to LLaMA-1 and other documented CommonCrawl-based models.

4.3. European Heritage Digitisation Progress (ENUMERATE 2012–2017)

The core survey conclusions of the four ENUMERATE Core Surveys are summarised in **Table 4** and are based on published survey reports [14].

Table 4. ENUMERATE Core Survey Results: Digitisation of European Cultural Heritage Institutions (2012–2017).

ENUMERATE Metric	Core Survey 1 (2012)	Core Survey 2 (2014)	Core Survey 3 (2015)	Core Survey 4 (2017)	Interpretation for DH
Institutions with a digital collection	83%	87%	84%	82%	High and stable presence: 4 out of 5 European institutions have some digital collection, though depth of coverage within collections varies widely.
% of collections digitised	20%	17%	23%	16%	Persistent structural gap: fewer than 1 in 4 objects digitised even in Europe's best-funded heritage sector. The 2017 dip to 16% reflects a larger and more diverse respondent pool.
% of collections still needing digitisation	57%	52%	50%	~40%	Despite a decade of EU-funded effort, a large majority of European analogue heritage remains undigitised. The
Institutions with a written digital strategy	34%	36%	34%	42%	Modest but steady improvement. As of 2017, the majority of European heritage institutions still lacked a formal digitisation policy, the baseline for any systematic global programme.
Institutions monitoring online use of digital collections	42%	51%	N/A	N/A	Not asked in Core Surveys 3 and 4. The indicator reflects limited investment in access analytics even among well-resourced European institutions.
Sample coverage	~1,000 institutions, 29 countries	~1,371 institutions	~1,030 institutions, 28 countries	~1,000 institutions, 28 countries	All four surveys cover European cultural heritage institutions only (museums, libraries, archives, other). These are unweighted institution-level figures; the ENUMERATE reports note this may underestimate actual digitised volume.

Source: All figures directly extracted from ENUMERATE/Europeana Core Survey reports. Core Survey 1 (2012): Survey Report on Digitisation in European Cultural Heritage Institutions 2012. Core Survey 2 (2014): ENUMERATE Digitisation Survey 2014. Core Survey 3 (2015): Survey Report 2015. Core Survey 4 institutional presence and strategy figures (84%, 42%); Europeana Pro indicator page.

The ENUMERATE data show an ongoing pattern of a structural deficit. Most European heritage institutions (83–87%) share that at least part of their collection can be digitised, but only a small percentage of collections (on average less than 25% during four rounds of the questionnaire over 5 years) are fully digitised. The lowest early figure, 17%, was recorded in the 2014 round (Core Survey 2), with a further dip to 16% in 2017. The spontaneous commentary shared by the ENUMERATE team pointed out that with such digitisation speed it will take at least 30 years for them to reach the digitisation needs of European institutions. It is the minimum requirement for the most well-furnished, most policy-supported and most technologically advanced heritage sector in the world.

What this means for global practice is that following 10 years and the dedicated EU funding in Europe, only 16–23% of the collections of Europe's institutions are digitised, and surely not as much in sub-Saharan Africa, South Asia or the Amazon basin. This asymmetry has been unquantifiable so far, as no ENUMERATE-like survey has been conducted on a global scale, but it is evident that it does exist.

5. The Epistemological Asymmetry Framework (EAF)

Combined, the three datasets examined in Sections 4.1–4.3 help to bring to light the epistemological problem of digital humanities, which is larger than each particular finding, namely, how Western, English-centred knowledge communities are mirrored in the training corpora of LLMs, in collections that are digitised for their training, and in the institutions that can deploy them. The Epistemological Asymmetry Framework (EAF) is suggested here as a conceptual structure of the diagnosis translated into actionable dimensions for DH researchers and institutions.

The EAF supports only the dissemination of the three main datasets that are the basis for this study; namely, LLaMA-1 pre-training composition data [4], the W3Techs content language statistics [1] and the ENUMERATE Core Survey series [14]. Every dimension of the EAF addresses an analytical question that is answerable (in whole or in part) by numbers that are immediately available from these sources. This methodological limitation is not accidental: The process of auditing must rely on data that can be verified or interpolated, and that data must be abundantly and clearly available. For data that are missing (especially where they cannot be obtained for data from the 'Global South'), this is indicated as an 'evidential finding' in order to achieve absolute transparency.

5.1. Dimension I: Linguistic Coverage in Training Data

Perhaps the most directly verifiable dimension of epistemological asymmetry is the language composition of the corpora that are used as training sets for LLM models. The amount of explicitly English-filtered data used for training LLaMA-1 is reported herein, directly from **Table 1** of Touvron et al. (2023), to be 82% [4]. This prolif-

eration of English-language domination is not only a result of the present distribution of web content; English is used to make up about 49–56% of all content in top websites, adding 26–33% points of English bias to the already-imbalanced web.

For researchers, Dimension I would correspond to an audit question; Is the LLM tool deployed in a concrete task (project) sufficiently well trained with the languages included in the selected corpus? If it is about LLMs for Swahili oral-history transcriptions, Quechua archival documents or Tamil literary texts, the answer, per the documented corpus compositions of LLaMA-1 and other CommonCrawl-based models with published training tables (e.g., T5), is nearly always no; for models without published corpus composition (e.g., GPT-4), the absence of disclosure itself precludes any affirmative answer. As a recommended intervention, multilingual corpus auditing should precede deployment, along with collaboration with low-resource language initiatives like CLARIN, DOBES (Documentation of Endangered Languages) and the Endangered Languages Documentation Programme.

5.2. Dimension II: Heritage Collection Digitisation

The second dimension relates to the archival substrate that can, in principle, be learnt from by LLMs, and, in practice, analysed by researchers. According to the four rounds of ENUMERATE Core Survey (2012–2017) reported in **Table 4** (in Section 4.3), even the world's best resourced and technically most advanced cultural heritage institutions reported having digitised only a small portion of their own collections (16–23%). The last systematic data Core Survey 4 (2017) showed 16% of analogue collections being reproduced digitally with 40% outstanding in terms of collecting.

The main result across this dimension is the absence of any comparable ENUMERATE-like survey of heritage from non-European regions, and therefore a lack of any empirical measure on the extent to which cultural heritage in Sub-Saharan Africa, South Asia, Latin America and the Pacific has been digitised. This data gap is not merely neutral, it is one part of structural inequalities that first caused the under-digitisation. It is recommended that there be two approaches in intervening by supporting ongoing pilot programmes (in Libya, Jamaica and Sri Lanka with UNESCO IFAP, for example) as examples in which scaled digitisation can be trialled, and advocating for the development of digitisation survey infrastructure in non-European countries that equals ENUMERATE.

5.3. Dimension III: Institutional Readiness and Access

The third dimension joins the previous two (even in countries where adequate training-data coverage and digitisation of collections are available) but access to these tools in the right hands can be hindered by the lack of institutional access for researchers and practitioners in low-income settings. The most reliable proxy is the ENUMERATE Core Survey 4 (2017) where 42% of European heritage institutions currently have a digital strategy in place (up from 34% in 2015) [14]. This is the highest that is known of the world's most institutionally developed heritage sector and reveals that, even within Europe, a lot of institutions are without a formal structure for systematic engagement through digital means.

Equivalent data is lacking outside Europe. These institutional constraints in Global South countries, though, are well documented in the wider digital-divide literature, including bandwidth restrictions, hardware costs, proprietary API charges, English-only documentation, and the absence of local technical support. Developing offline capabilities and providing open-access API tiers to low-income institutions and enforcing formal digitisation policy as a ground rule for LLM-assisted DH investments are thus the suggested solutions for this dimension.

The three components of the EAF represent the epistemological asymmetry (from the upstream source of training data composition, to the infrastructural substrate of digitisation rates, and then to the downstream condition of possibility of institutional access). Verification of at least some information that is publicly available makes measurement of each dimension possible [21]. Recommendations in each dimension are broad, specific, and actionable. Infrastructure investments are needed for equitable practice to be measurable as the absence of digitisation surveys in the Global South is just as much a call for infrastructure investment as are the gaps in other dimensions.

6. Discussion

The present study employed three sets of data that are consistent and reinforcing in terms of their depictions of epistemological asymmetry; as noted in Section 3.2, they are treated as complementary diagnostic dimensions

rather than as causally linked. It is easy to confirm that 82% of the data used to train LLaMA-1 has actually been explicitly English-filtered, as can be seen from **Table 1** of Touvron et al. (2023) [4]. According to W3Techs data, English accounts for approximately 49–56% of top-website content [1]; while according to Internet World Stats (2020), approximately 25.9% of worldwide internet users use English as their main language [7]. The increase from 25.9% (internet users) to 82% (training data) is not a statistical artifact, but rather the result of design decisions, each of which is documented and citable.

The ENUMERATE data further complicate this picture. After ten years of dedicated investment, Europe's best-funded heritage sector still has 16–23% of its collections digitised and if there were a comparable index for any other region, the set of images on which LLMs are trained would not just be underrepresented in terms of non-Western languages, it would be, structurally, an archive of the already-advantaged [22,23]. What emerged from this training input is the findings of Tao et al. (2024) all clustering together in English-speaking Protestant Europe in terms of cultural values, which makes the dataset an overwhelmingly Western, predominantly English dataset of digital traces left behind by well-resourced institutions and individuals [24].

There are some constraints in this analysis, primarily due to focusing on LLaMA-1 as the key demonstrative LLM case study. The subsequent models, LLaMA-2, LLaMA-3, GPT-4 and Gemini, do not have comparable training composition tables published, so their corpora are much less precisely verifiable. To adhere to the principle of verifiability central to this paper, LLaMA-1 was selected because its training data can be cited directly. Where LLM developers do not disclose the composition of their training datasets, this type of audit becomes impossible; the absence of disclosure is itself a barrier to scholarly accountability. Future work and advocacy for mandatory training-data disclosure would address this gap directly.

The other area of limitation is the time difference in ENUMERATE data due to the fact that the latest Core Survey was in 2017. Since then digitisation has progressed and the percentage of Europe that is digitised is probably even greater now. The ENUMERATE Observatory has observed the planning of a new survey round. However, the most important result, that there was no equivalent global survey, does not lose its validity. Without systematic and regular digitisation surveys as is carried out by ENUMERATE in Europe, the degree of global asymmetry can never be quantified until digitisation is carried out with the same devotion and regularity in the Global South.

7. Conclusion

Based solely on authenticated and directly citable primary-source data, this study demonstrates that LLaMA-1 (as a representative and well-documented exemplar of CommonCrawl-based LLM training regimes) introduces a substantial epistemological asymmetry. The LLaMA-1 training data contains 82% English-filtered data. Top website content is about 49–56% English language while approximately 25.9% of the worldwide internet audience uses English as their primary language [7]. Only four ENUMERATE survey rounds have recorded the digitisation of collections in European cultural heritage institutions, and then only 16–23% have done so, compared to other regions where fewer institutions with fewer resources have no equivalent survey at all. All five GPT models evaluated [24] cluster in the cultural values of English-speaking Protestant Europe, with cultural prompting failing for 19–29% of countries tested.

These findings are not speculative in any way as each of the numbers used in this paper is quoted on an immediately extractable basis from the primary ones cited in the references. Not only is it a methodological but also a scholarly devotion. If the data is not extracted but fabricated, the epistemological asymmetries in DH practice cannot be measured or addressed. This paper illustrates the discipline of evidence-based critique.

The Epistemological Asymmetry Framework (EAF) introduced here offers researchers and institutions an auditable instrument developed and designed based on verifiable indicators. Perhaps its greatest contribution could be in the systematic measurement of heritage digitisation outside Europe. Without this measurement, the field is working in a context it can't fully understand; and with the context of inequity, the first place to start is the question of what we don't know.

Further areas of enquiry for future studies include: (1) devising ENUMERATE-equivalent digitisation surveys for non-European areas; (2) promoting mandatory training data disclosure for the composition of training sets for LLM developers; and (3) developing cultural alignment audits for models other than the GPT family of models, given the growing popularity of open-source models with published corpus compositions, such as Mistral, Falcon and BLOOM, in DH contexts.

Author Contributions

Conceptualization, M.P. and C.P.R.; Methodology, M.P. and C.P.R.; Software, M.P.; Validation, M.P. and C.P.R.; Formal Analysis, M.P.; Investigation, M.P. and C.P.R.; Resources, M.P. and C.P.R.; Data Curation, M.P.; Writing—Original Draft Preparation, M.P.; Writing—Review and Editing, M.P. and C.P.R.; Visualization, M.P.; Supervision, C.P.R.; Project Administration, C.P.R. Both authors have read and agreed to the published version of the manuscript.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

The study was conducted in accordance with the Declaration of Helsinki. Formal institutional ethics approval was obtained from the Research and Development Department and the School of Media Studies at Presidency University in Bengaluru, India (January 2026).

Informed Consent Statement

Not applicable.

Data Availability Statement

All datasets used in this study are publicly accessible from the sources cited. No primary data were collected by the authors.

Conflicts of Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

AI Use Statement

The authors used Grammarly for proofreading and language correction only. The substantive content of this paper (including the conceptual framework, and all scholarly arguments) was authored entirely by the researchers. The authors reviewed all content and take full responsibility for the integrity of the published work.

References

1. W3Techs. Historical Trends in the Usage Statistics of Content Languages for Websites. Available online: https://w3techs.com/technologies/history_overview/content_language/ms (accessed on 30 April 2026).
2. Bender, E.; McMillan-Major, A.; Shmitchell, S.; et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Online, 3–10 March 2021; pp. 610–623. [CrossRef]
3. Blodgett, S.L.; Barocas, S.; Daumé III, H.; et al. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 5454–5476. [CrossRef]
4. Touvron, H.; Lavril, T.; Izacard, G.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint* **2023**, *arXiv:2302.13971*. [CrossRef]
5. Dodge, J.; Sap, M.; Marasović, A.; et al. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 1286–1305. [CrossRef]
6. Luccioni, A.; Viviano, J. What's in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; pp. 182–189. [CrossRef]
7. Internet World Stats. *Internet World Users by Language: Top 10 Languages Used in the Web*; Internet World Stats: Bogotá, Colombia, 2020.

8. OBDILCI. English@Web: An Historical Bias. Available online: https://www.obdilci.org/projects/main/englishweb/ (accessed on 8 May 2026).
9. Baack, S. *Training Data for the Price of a Sandwich: Common Crawl's Impact on Generative AI*; Mozilla Foundation: San Francisco, CA, USA, 2024. Available online: https://assets.mofoprod.net/network/documents/Common_Crawl_Mozilla_Foundation_2024.pdf
10. Raffel, C.; Shazeer, N.; Roberts, A.; et al. Exploring the Limits of Transfer Learning with a Unified Text-To-Text Transformer. *arXiv preprint* **2019**, *arXiv:1910.10683*. [CrossRef]
11. Joshi, P.; Santy, S.; Budhiraja, A.; et al. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 6282–6293. [CrossRef]
12. Ahuja, K.; Diddee, H.; Hada, R.; et al. MEGA: Multilingual Evaluation of Generative AI. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 4232–4267. [CrossRef]
13. Lai, V.D.; Ngo, N.; Veyseh, A.P.B.; et al. ChatGPT beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 6–10 December 2023; pp. 13171–13189. [CrossRef]
14. Europeana Pro. ENUMERATE Results. Available online: https://pro.europeana.eu/page/results (accessed on 30 April 2026).
15. Sheldon, Z. The Archived Web: Doing History in the Digital Age. *Inf. Commun. Soc.* **2019**, *23*, 309–311. [CrossRef]
16. UNESCO. *Charter on the Preservation of the Digital Heritage*; UNESCO: Paris, France, 2009. Available online: https://unesdoc.unesco.org/ark:/48223/pf0000179529.locale=en
17. UNESCO. Advancing Access to Information and Digital Preservation: IFAP's Impact on Digitizing Documentary Heritage. Available online: https://www.unesco.org/en/articles/advancing-access-information-and-digital-preservation-ifaps-impact-digitizing-documentary-heritage (accessed on 28 April 2026).
18. Risam, R. *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy*; Northwestern University Press: Evanston, IL, USA, 2018.
19. Noble, S.U. *Algorithms of Oppression: How Search Engines Reinforce Racism*; New York University Press: New York, NY, USA, 2018.
20. Couldry, N.; Mejias, U.A. Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject. *Televis. New Media* **2019**, *20*, 336–349.
21. Fiormonte, D.; Chaudhuri, S.; Ricaurte, P. *Global Debates in the Digital Humanities*; University of Minnesota Press: Minneapolis, MN, USA, 2022.
22. Cao, Y.; Li, Z.; Lee, S.; et al. Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. In Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), Dubrovnik, Croatia, 5 May 2023; pp. 53–67. [CrossRef]
23. Trott, S. Large Language Models and the Wisdom of Small Crowds. *Open Mind* **2024**, *8*, 723–738. [CrossRef]
24. Tao, Y.; Viberg, O.; Baker, R.S.; et al. Cultural Bias and Cultural Alignment of Large Language Models. *PNAS Nexus* **2024**, *3*, e346. [CrossRef]