Whose Data, Whose Benchmark?
A User-Centric Framework for Auditing Evaluation Resources

Mir Tafseer Nayeem
University of Alberta
mnayeem@ualberta.ca

Davood Rafiei
University of Alberta
drafiei@ualberta.ca

arXiv:2610.19268v1 [cs.CL] 13 Oct 2026

Abstract

Benchmarks shape how model capabilities are measured and interpreted, yet their scores can obscure whose conventions, contexts, and judgments define successful performance. We introduce a *user-centric audit framework* that distinguishes users who contribute benchmark content from those whose needs evaluation results are intended to represent. The framework operationalizes seven dimensions across benchmark content, metadata, documentation, and scoring procedures. Across 16 resources, we find that American-centered conventions account for 62.4–94.4% of matched signals in 11 text benchmarks, while geographic visibility in five image–text resources is concentrated and strongly associated with national GDP share (mean Pearson $r = 0.780$; Spearman $\rho = 0.778$). Documentation reveals a systematic source–target gap: all 16 resources operationalize source-user context, but none operationalize target-user context. Paired scoring experiments further show that meaning-preserving answer variants reduce DROP exact-match accuracy from 100.0% to 33.3%, while TruthfulQA exhibits metric-dependent reductions in reference similarity. Our framework and documentation checklist make user assumptions auditable, clarifying what benchmark scores represent and where evaluation claims are supported.[1]

1 Introduction

Large Language Models (LLMs) and multimodal AI systems are increasingly used across education, healthcare, information access, and everyday decision-making. Their users differ in language, cultural conventions, expertise, geographic context, and expectations. As these systems reach broader populations, benchmarks play a growing role in defining how capabilities are measured, compared, and communicated. Recent research has examined benchmark validity, evaluation reliability, and dataset documentation (Raji et al., 2021; Liu et al., 2024; Mousavi et al., 2026). Yet a fundamental question remains: *whose users are represented in the resources that define successful performance?*

Language and multimodal data are produced by people, interpreted by people, and deployed for people. Evaluation resources consequently inherit assumptions about the people and contexts from which their content originates. A benchmark may favor particular spelling, date, or measurement conventions; image–text collections may disproportionately reference certain countries; and accepted answers may depend on annotator judgments or scoring rules. Prior work has highlighted the importance of documenting data provenance and linguistic diversity (Bender and Friedman, 2018; Gebru et al., 2021; Faisal et al., 2024). However, understanding how these assumptions collectively shape the scope and interpretation of benchmark claims remains an important challenge.

We distinguish between *source users*, whose content and judgments contribute to evaluation resources, and *target users*, whose needs and contexts the resulting scores are intended to inform. Documenting where benchmark content originates does not establish which user populations or settings its scores represent. This distinction motivates a *user-centric audit*: a systematic examination of the assumptions embedded in evaluation resources and their implications for what benchmark scores measure and where evaluation claims apply. Such an audit considers whose conventions are treated as defaults, which geographies are visible, whose judgments define accepted answers, and whether intended evaluation contexts are documented.

We introduce a *user-centric framework for auditing evaluation resources* that operationalizes these questions across *seven* dimensions: linguistic and locale conventions, geographic visibility, socioeconomic structure, source-user context, target-user assumptions, annotator visibility, and evaluation nor-

[1] Data, code, and research artifacts are publicly available at tafseer-nayeem.github.io/WhoseBench.

malization. By connecting evidence from benchmark content, metadata, documentation, and scoring procedures, the framework provides a unified approach to examining resource-level assumptions and assessing the scope of evaluation claims.

We instantiate the framework through three empirical audits spanning 16 resources. First, we examine American-centered linguistic and locale conventions across 11 text benchmarks, including how meaning-preserving answer variants affect scoring outcomes. Second, we measure country visibility in captions and associated metadata from five image–text resources and examine its relationship with socioeconomic structure. Third, we assess whether documentation across all 16 resources makes source users, target contexts, annotator information, locale assumptions, and evaluation procedures explicit and operationally inspectable. These complementary analyses connect benchmark composition, evaluation rules, and documented user contexts within a common framework.

Our contributions are threefold:

1. We introduce a reusable *user-centric audit framework* that distinguishes source users from target users and operationalizes seven dimensions of user-relevant assumptions across evaluation resources (§3).

2. We systematically audit 16 text and image–text resources, revealing American-centered convention asymmetries (§4), geographic concentration associated with socioeconomic structure (§5), and a persistent source–target documentation gap (§6). Paired scoring experiments further demonstrate how valid answer conventions can affect evaluation outcomes.

3. We provide an actionable documentation checklist and reproducible audit artifacts that support benchmark creators and users in examining resource assumptions and interpreting the scope of evaluation claims (§7).

2 Related Work

Benchmark validity and documentation. Prior research establishes how benchmark design and construction shape the measurement and interpretation of model capabilities (Paullada et al., 2021; Raji et al., 2021; Miller, 2022; Liu et al., 2024), including the influence of benchmark content and scoring procedures on evaluation reliability (Mousavi et al., 2026). Documentation frameworks, including data statements, datasheets, and BenchmarkCards, provide mechanisms for reporting provenance, intended use, and evaluation conditions (Bender and Friedman, 2018; Gebru et al., 2021; Sokol et al., 2025). We operationalize these concerns through a user-centric audit framework that connects resource composition, scoring conventions, and documented user contexts to the validity and scope of benchmark evaluation claims.

User-centric and linguistic evaluation. Research on user-centric evaluation addresses real-world user intents (Wang et al., 2024a), performance across dialects and language varieties (Faisal et al., 2024), cultural variation in time expressions (Shwartz, 2022), and American English preferences across the LLM pipeline (Nayeem and Rafiei, 2026). Population-specific data construction for children's language (Nayeem and Rafiei, 2024) and personalized opinion summarization with context-sensitive evaluation (Nayeem and Rafiei, 2025) further demonstrate the importance of aligning resources and evaluation procedures with intended users. We extend these insights to the *evaluation resources* themselves, systematically measuring linguistic and locale conventions, testing how scoring procedures handle meaning-preserving alternatives, and examining whether documented target-user assumptions support broader evaluation claims.

Geographic representation in multimodal resources. Prior studies identify uneven geographic coverage in visual datasets and regional disparities in model performance (Shankar et al., 2017; DeVries et al., 2019). More recently, Basu et al. (2026) analyzed country representation in image–text datasets through caption-based geographic profiling. We integrate geographic visibility and its socioeconomic associations into the user-centric audit, connecting patterns of country representation with source provenance, intended user contexts, and the interpretation of evaluation results.

3 A User-Centric Framework for Auditing Evaluation Resources

A user-centric audit begins with a premise: *language and multimodal data are produced by people, interpreted by people, and deployed for people*. Evaluation resources therefore reflect choices about whose language, knowledge, experiences, and judgments are represented. Our framework systemati-

Audit dimension	User-centric question	Operational signals in this paper
Linguistic and locale conventions	Whose spelling, formatting, and measurement conventions are treated as default?	Spelling variants (e.g., color/colour), date formats, time notation (e.g., 7 pm/19:00), and measurement units (e.g., miles/kilometres) in benchmark questions, passages, answer options, and gold answers.
Geographic visibility	Which countries and regions are visible in image–text resource descriptions?	Explicit country references in captions and textual metadata, resolved using country names, demonyms (e.g., Japanese), capitals, and selected cities (e.g., Paris).
Socioeconomic structure	How does geographic visibility relate to economic and socioeconomic indicators?	Associations with GDP share, population share, GDP per capita, Internet usage, and tourism arrivals.
Source-user context	Where does benchmark content originate, and whose contexts shape its collection?	Documented sources (e.g., examinations, Wikipedia, Reddit, CommonCrawl) and provenance metadata, such as website domains (e.g., `bbc.co.uk`, `reddit.com`) and country-code TLDs (e.g., `.uk`, `.ca`, `.jp`).
Target-user assumptions	Which populations and settings are evaluation claims intended to represent?	Documented intended users (e.g., students, clinicians), evaluation audiences, deployment contexts, and conditions for interpreting benchmark results.
Annotator visibility	Whose judgments determine labels, references, and annotation quality?	Documented annotator expertise (e.g., medical professionals), demographics, locale (e.g., country, language variety), recruitment (e.g., crowdsourcing), compensation, annotation protocols, and quality-control procedures (e.g., inter-annotator agreement).
Evaluation normalization	Do scoring procedures recognize meaning-preserving answer alternatives?	Documented normalization rules, accepted variants (e.g., color/colour), locale conventions, and paired scorer-sensitivity tests using valid answer alternatives.

Table 1: User-centric audit taxonomy for evaluation resources. Seven dimensions connect user-relevant assumptions to measurable and inspectable evidence across benchmark content, metadata, documentation, and scoring procedures.

cally examines these choices, making the assumptions embedded in benchmark content, metadata, documentation, and scoring procedures explicit and auditable. This provides a basis for understanding what benchmark scores represent and the user contexts in which their interpretation is supported.

Source users and target users. We distinguish *source users*, whose content and judgments contribute to an evaluation resource, from *target users*, whose needs and contexts its results are intended to inform. Source users include exam writers, crowdworkers, and web contributors, while target users may include students, clinicians, and multilingual speakers. Consider a medical benchmark constructed from professional examination questions and subsequently used to evaluate systems intended for clinicians across healthcare settings. Its documentation may establish where the questions originated without specifying how the resulting scores apply across those settings. The source–target distinction thus connects benchmark provenance to the scope of its evaluation claims.

Dimensions of a user-centric audit. Table 1 organizes the framework into *seven* dimensions: linguistic and locale conventions, geographic visibility, socioeconomic structure, source-user context, target-user assumptions, annotator visibility, and evaluation normalization. Each dimension translates a question about users into measurable or inspectable evidence drawn from benchmark content, metadata, documentation, or evaluation procedures. The resulting framework connects how a resource is constructed, whose contexts it represents, and how its scores are determined.

From resource assumptions to evaluation claims. Benchmark composition and scoring rules shape the evidence used to assess model capabilities. A meaning-preserving answer expressed as *colour*, *kilometres*, or a day–month date may receive a different score when the evaluator recognizes only another convention. Likewise, concentrated geographic references in image–text captions and metadata make some contexts more visible than others. These examples illustrate how linguistic defaults, geographic representation, and evaluation normalization influence the interpretation of benchmark results. Auditing these properties makes the underlying assumptions available for systematic examination before evaluation scores are generalized across users and settings.

We apply the framework through three complementary empirical audits: linguistic and locale conventions in text benchmarks (§4), geographic visibility and socioeconomic structure in image–text resources (§5), and documentation of user-relevant assumptions across all resources (§6).

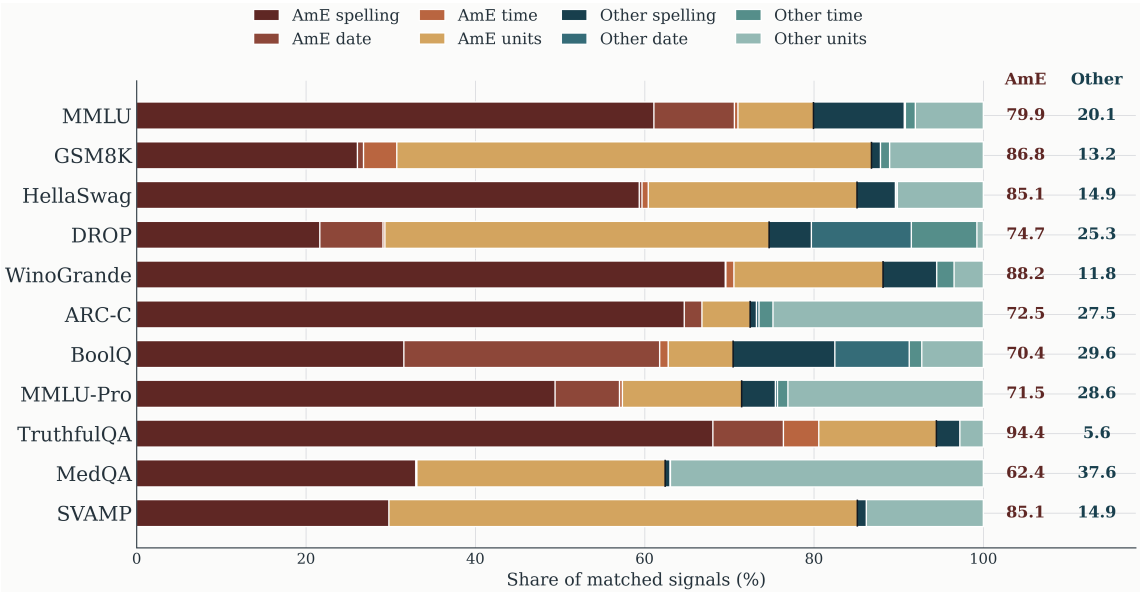

Figure 1: Linguistic and locale convention signals across 11 text benchmarks. Bars compare American-centered and other user-valid matches in spelling, dates, time, and measurement units; rightmost columns show aggregate shares. American-centered conventions predominate, with variation across benchmarks and convention types. Appendix Figure 4 provides the breakdown.

A paired scoring-sensitivity analysis additionally tests how evaluation procedures respond to meaning-preserving answer variants (§4.2). This integrated approach connects the assumptions present in evaluation resources to the evidence and scoring decisions underlying benchmark claims.

4 RQ1: Do Text Benchmarks Encode an Implicit Default User?

Text benchmarks assess knowledge, reasoning, and comprehension through questions and reference answers that also encode linguistic and locale conventions. We audit which conventions these resources favor and how valid differences in answer expression affect evaluation scores.

Benchmarks and audit setup. We examine 11 benchmarks spanning knowledge, mathematics, commonsense reasoning, reading comprehension, science, truthfulness, and medical QA: MMLU, GSM8K, HellaSwag, DROP, WinoGrande, ARC-Challenge, BoolQ, MMLU-Pro, TruthfulQA, MedQA, and SVAMP. Several feature prominently in model evaluations (Grattafiori et al., 2024; Team et al., 2025). Appendix Table 7 summarizes their tasks and formats.

We audit four convention classes: spelling, date formats, time notation, and measurement units. English varieties have developed through distinct historical and sociolinguistic contexts, including the global spread of English and postcolonial language practices (Crystal, 2003; Tikly, 2016). Some conventions are shared across regions, while others may differ. We group selected alternatives under *other user-valid conventions* for comparative measurement, recognizing that these forms do not constitute a single English variety. Appendix A provides the convention inventory and detection rules.

Method. We detect American-centered and other user-valid convention signals in benchmark questions, passages, instructions, answer options, and gold answers. Each item is classified as American-centered only, other conventions only, mixed, or unmatched. We report *coverage*, the percentage of items containing at least one audited signal, and *American-centered share*, the percentage of matched signals belonging to the American-centered category. We also exclude spelling matches to assess whether the asymmetry extends to date, time, and measurement conventions. Appendix A details the aggregation procedure.

4.1 Results and Analysis

American-centered conventions dominate matched signals. Figure 1 shows an American-centered majority in all 11 benchmarks, ranging from 62.4% in MedQA to 94.4% in TruthfulQA. The pattern spans general knowledge, mathematics, commonsense reasoning, and specialized QA.

Coverage varies substantially across resources. Table 2 shows that audited conventions occur in 5.97% of WinoGrande items and 7.59% of Truth-

Benchmark	Coverage (%)	AmE excl. spelling (%)
MMLU (Hendrycks et al., 2021)	26.60	66.9
GSM8K (Cobbe et al., 2021)	25.40	83.3
HellaSwag (Zellers et al., 2019)	38.11	71.2
DROP (Dua et al., 2019)	78.08	72.3
WinoGrande (Sakaguchi et al., 2021)	5.97	77.2
ARC-Challenge (Clark et al., 2018)	18.73	22.5
BoolQ (Clark et al., 2019)	44.92	68.9
MMLU-Pro (Wang et al., 2024c)	37.80	47.3
TruthfulQA (Lin et al., 2022)	7.59	90.5
MedQA (Jin et al., 2021)	65.85	44.3
SVAMP (Patel et al., 2021)	9.30	80.0

Table 2: Convention coverage and spelling-excluded robustness across 11 text benchmarks. Coverage is the share of items containing an audited signal; the final column reports the American-centered share among non-spelling matches.

fulQA items, compared with 65.85% in MedQA and 78.08% in DROP. American-centered share captures the distribution of *matched signals*, while coverage measures their prevalence across benchmark items. Reporting both distinguishes convention prevalence from which forms predominate.

The asymmetry extends beyond spelling. Excluding spelling matches leaves an American-centered majority in eight of 11 benchmarks. The highest shares occur in TruthfulQA (90.5%), GSM8K (83.3%), SVAMP (80.0%), and DROP (72.3%). ARC-Challenge (22.5%), MedQA (44.3%), and MMLU-Pro (47.3%) show the opposite pattern. These results reveal convention profiles across benchmarks and establish that the asymmetry extends beyond orthographic variation.

4.2 Scorer Case Study: Do Valid Answer Variants Change Evaluation Scores?

The convention audit identifies which forms appear in benchmark content. We next test whether *meaning-preserving changes in answer conventions alter evaluation outcomes* using controlled answer pairs and scoring procedures. This resource-level analysis isolates scoring sensitivity from *model-specific* behavior, establishing how evaluation rules recognize valid answer alternatives.

Paired scoring experiment. We select DROP and TruthfulQA for their released answer-text evaluation procedures. For DROP, we use the official exact-match (EM) and token-level F1 scorer (Dua et al., 2019); for TruthfulQA, we reproduce the released BLEU/ROUGE correct-versus-incorrect reference-ranking procedure (Lin et al., 2022). Other benchmarks, such as MMLU and ARC-Challenge, primarily score multiple-choice labels, while GSM8K evaluates extracted numerical answers. Prior research on multiple-choice evaluation and free-form answer matching further motivates directly testing how scoring procedures handle valid answer alternatives (Zheng et al., 2024; Wang et al., 2024b; Chandak et al., 2025).

Benchmark	Metric	Orig.	Var.	Δ [95% CI]
DROP	EM	100.0	33.3	-66.7 $[-92.3, -36.4]$
DROP	F1	100.0	77.2	-22.8 $[-41.6, -5.2]$
TQA	BLEU correct.	97.0	97.0	0.0 $[0.0, 0.0]$
TQA	ROUGE-1 correct.	97.0	97.0	0.0 $[0.0, 0.0]$
TQA	ROUGE-2 correct.	100.0	93.9	-6.1 $[-16.1, 0.0]$
TQA	ROUGE-1 sim.	100.0	88.9	-11.1 $[-13.5, -9.0]$

Table 3: Scorer sensitivity to answer variants. DROP uses official EM/F1; TQA (TruthfulQA) uses BLEU/ROUGE reference-ranking rules. Changes are in percentage points for EM, F1, and correctness, and on a 0–100 scale for similarity. CIs use question-clustered bootstrap resampling.

For each eligible pair, we compare an original reference-consistent answer with a meaning-preserving spelling, written-date, or time-format variant against the same gold annotations. We retain each evaluation procedure and estimate paired changes with 95% confidence intervals using question-clustered bootstrapping. Appendix A.6 details the transformations, screening, sample sizes, scoring details, and uncertainty estimates.

DROP: Valid alternatives can lose substantial credit. Table 3 shows that DROP's official EM decreases from 100.0% to 33.3% for meaning-preserving variants, a reduction of 66.7 percentage points (95% CI $[-92.3, -36.4]$). Token-level F1 also decreases from 100.0% to 77.2% (-22.8 points; 95% CI $[-41.6, -5.2]$). Partial credit reduces the penalty, but valid changes in answer expression still affect evaluation scores.

TruthfulQA: Lower similarity need not change correctness. TruthfulQA responds differently. BLEU and ROUGE-1 correctness remain at 97.0%, even as ROUGE-1 similarity to a correct reference decreases from 100.0 to 88.9 (95% CI for the change $[-13.5, -9.0]$). ROUGE-2 correctness declines by 6.1 points, with a confidence interval that includes zero. Valid variants can reduce reference similarity while preserving the evaluator's correctness decision. The contrast with DROP demonstrates how scoring rules determine the consequences of differences in answer expression.

Implications for user-aware evaluation. These findings distinguish *conventions represented in benchmark content* from *alternatives recognized by scoring procedures*. Exact match, token over-

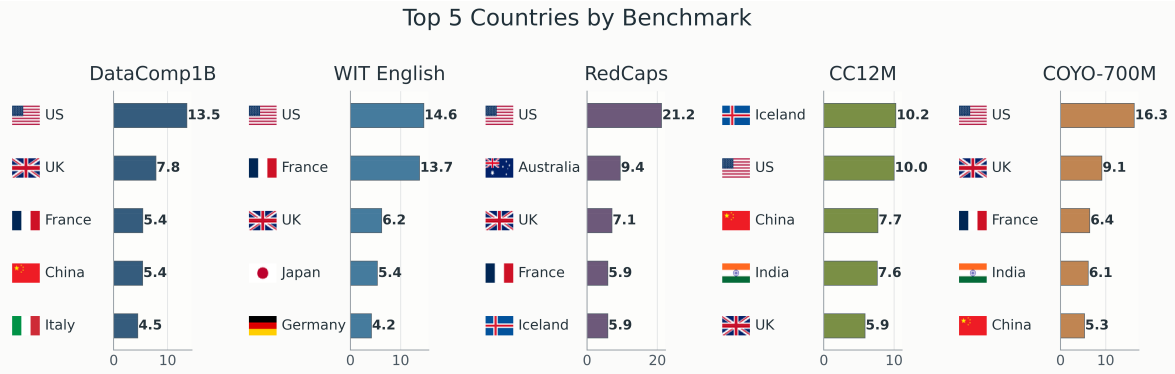

Figure 2: **Caption geography is concentrated in a few countries.** Bars show the top five country shares from caption-derived assignments in each resource. The US leads four resources, while Iceland leads CC12M, revealing source-specific variation. Countries are identified through deterministic text matching, without image geolocation.

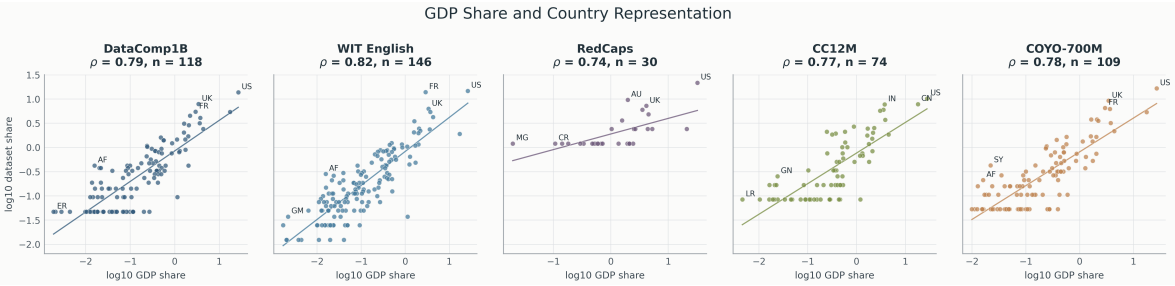

Figure 3: **Country visibility is strongly associated with GDP share.** Each point represents a country with at least one million people. GDP and caption-derived country shares are normalized over the same retained countries and plotted on log scales. Fitted trends and Spearman correlations summarize the association across resources.

lap, and reference ranking can score meaning-preserving responses differently. Benchmark creators can test consistency across spelling, date, time, and measurement variants and document normalization rules. Such audits reveal how answer conventions influence evaluation outcomes and inform the user-centric documentation checklist (§7).

5 RQ2: Which Geographies Are Visible in Multimodal Benchmarks?

Image–text resources reflect geographic contexts through captions and associated metadata, shaping which countries become explicitly visible in multimodal data. We audit the distribution of these geographic references across five resources and examine their relationship with national socioeconomic indicators and language composition.

Resources and audit setup. We analyze five image–text resources with distinct collection sources: DataComp1B, WIT English, RedCaps, CC12M, and COYO-700M. These span web-scale image–text collections, Wikipedia metadata, Reddit captions, web alt-text, and CommonCrawl-derived pairs, respectively (Appendix Table 10).

We identify explicit geographic references in captions and associated textual metadata using a deterministic resolver covering country names, aliases, demonyms, capitals, and selected cities. Records are assigned to countries when the textual evidence yields a unique resolution. We define *geographic visibility* as the distribution of country assignments among resolved records, measuring which geographies appear explicitly in each resource's textual layer. Appendix B details the lexicon, matching rules, and assignment procedure.

5.1 Results and Analysis

Geographic visibility is concentrated across sources. Figure 2 shows that the United States leads four resources: DataComp1B (13.55%), WIT English (14.59%), RedCaps (21.18%), and COYO-700M (16.30%). CC12M differs, with Iceland first (10.24%) and the United States second (9.99%). Across resources, the top three countries account for 26.8–37.6% of resolved assignments, while Europe, North America, and Oceania collectively account for 60.9–76.5%. These patterns reveal geographic concentration across distinct collection sources. Appendix B.4 reports the

Criterion	Score	Audit question
Source-user context	0–2	Does the documentation identify who produced the content, including source populations, platforms (e.g., Reddit, Wikipedia), institutions, or collection processes?
Target-user or evaluation context	0–2	Does it specify intended users (e.g., students, clinicians), evaluation audiences, deployment settings, or the contexts in which benchmark results are expected to apply?
Annotator visibility	0–2	Does it describe annotator expertise, demographics, locale, recruitment (e.g., crowdsourcing), compensation, quality control, or the absence of human annotation?
Locale assumptions	0–2	Does it document spelling (e.g., color/colour), date, time, measurement, currency, numeric-format, language-variety, or source-language conventions?
Evaluation normalization	0–2	Does it document how scoring procedures handle valid alternative answers, formatting conventions (e.g., date formats), or locale-specific equivalents?
Source/geographic metadata	0–2	Does it expose source URLs, website domains (e.g., `bbc.co.uk`, `reddit.com`), top-level domains (e.g., `.uk`, `.ca`, `.jp`), language tags (e.g., `en`, `fr`, `ja`), country or region fields, or other metadata supporting provenance and geographic auditing?

Table 4: Documentation audit rubric across six user-relevant criteria. Each criterion is scored as 0 (absent), 1 (mentioned explicitly), or 2 (operationalized through inspectable data, metadata, or evaluation procedures).

top-ten distributions and low-frequency countries.

Country visibility closely tracks economic scale. Figure 3 shows a strong association between country visibility and World Bank GDP share (mean Pearson $r = 0.780$; Spearman $\rho = 0.778$), with rank correlations ranging from $\rho = 0.735$ in RedCaps to $\rho = 0.823$ in WIT English. GDP share provides the strongest single-variable fit among socioeconomic indicators (mean $R^2 = 0.614$; count-weighted $R^2 = 0.748$). Population share, GDP per capita, Internet usage, and tourism arrivals yield weaker fits and modest improvements when combined with GDP share. These associations are descriptive rather than causal. Appendix B.6 details the regression specifications and comparisons.

Language composition changes visibility, but economic concentration persists. To assess language effects, we apply the same resolver to a balanced WIT sample spanning French, German, Spanish, and Japanese, motivated by documented concentration in English webtext (Dodge et al., 2021). In this separate comparison, the U.S. share falls from 20.0% in English to 8.5% in the non-English sample. Yet the ten largest economies account for 53.5% of resolved assignments, with strong GDP associations (Pearson $r = 0.752$; Spearman $\rho = 0.744$). Language composition shifts geographic visibility, while economic concentration persists (see Appendix B.7 for details).

Implications for user-aware evaluation. Geographic visibility varies with resource composition and language and is strongly associated with economic scale. Country references identify contexts visible in resource text without establishing the intended evaluation users. Documenting geographic coverage, languages, and source contexts makes these distinctions auditable and informs the user-centric documentation checklist (§7).

Criterion	Mean score	Ment. (%)	Oper. (%)
Source-user context	2.00	100.0	100.0
Target/eval. context	1.00	100.0	0.0
Annotator visibility	0.56	56.2	0.0
Locale assumptions	0.31	25.0	6.2
Evaluation normalization	1.44	75.0	68.8
Source/geographic metadata	0.69	37.5	31.2

Table 5: Documentation Audit results across 16 resources. Mean scores range from 0 (absent) to 2 (operationalized). "Ment." indicates resources scoring 1 or 2; "Oper." indicates those scoring 2, % are computed across all 16 resources.

6 RQ3: Are User Assumptions Made Explicit in Documentation?

Benchmark documentation defines how evaluation resources can be understood, inspected, and applied. Building on the linguistic and geographic audits in §4 and §5, we assess whether documentation makes six user-relevant dimensions explicit: source-user context, target-user context, annotator visibility, locale assumptions, evaluation normalization, and source/geographic metadata.

Resources and audit setup. We audit public documentation for all 16 resources: 11 text benchmarks and five image–text collections (Appendix Tables 7 and 10). Sources include benchmark papers, official dataset cards, and repositories, where available (Appendix Table 16).

Scoring protocol. Two annotators independently score each resource against the six criteria in Table 4: **0** (absent), **1** (mentioned), or **2** (operational-

Field	Minimum documentation question	Why it matters for evaluation
Source-user context	Who produced the content, and through which platform, institution, task, or collection process?	Identifies whose language, knowledge, experiences, and judgments are represented.
Target-user context	Which users, settings, or evaluation claims is the benchmark intended to represent?	Clarifies the populations and contexts to which benchmark results may apply.
Annotator context	Who provided labels or reference answers, and what is known about their expertise, locale, recruitment, compensation, and quality control?	Makes visible whose judgments determine correctness, relevance, or acceptability.
Locale conventions	Which spelling, date, time, unit, currency, numeric-format, or language-variety conventions appear in the data or scoring procedure?	Identifies regional assumptions that may influence benchmark content and evaluation outcomes.
Evaluation normalization	How does the scorer handle valid alternative answers, including spelling, formatting, measurement, and locale-specific equivalents?	Reveals whether equivalent responses receive consistent scores across user conventions.
Source/geographic metadata	Which URLs, website domains, top-level domains, language tags, or geographic fields are available for auditing?	Supports tracing content sources and examining geographic visibility without conflating source provenance with content location.

Table 6: Minimal user-centric documentation checklist for evaluation resources. The six fields translate the audit taxonomy into practical reporting questions that clarify benchmark provenance, intended users, and evaluation assumptions.

ized). A score of 1 requires an explicit statement; a score of 2 requires an inspectable mechanism, such as metadata fields, source partitions, annotation protocols, or evaluation rules. Each resource receives the highest score supported by documented evidence. Appendix C reports the annotation procedure and agreement.

6.1 Results and Analysis

Table 5 reports average scores and proportions of resources mentioning or operationalizing criteria.

Documentation identifies sources more clearly than intended users. All 16 resources operationalize source-user context through documented origins, including examinations, Wikipedia, Reddit, web alt-text, and CommonCrawl. Although all mention target-user or evaluation context, none operationalize it. This systematic gap shows that resource provenance is more clearly specified than the populations, settings, and needs that benchmark scores are intended to represent.

Annotators and locale conventions remain under-documented. Annotator information is mentioned in 56.2% of resources, with none operationalizing annotator attributes as auditable metadata. Locale assumptions receive even less attention: 25.0% mention them, and only 6.2% operationalize them. Consequently, the human judgments and linguistic conventions underlying evaluation references remain difficult to inspect.

Scoring rules are clearer than audit-enabling metadata. Evaluation normalization is mentioned in 75.0% of resources and operationalized in 68.8%. However, documented scoring procedures may still treat meaning-preserving answer variants differently, as demonstrated in §4.2. Source/-geographic metadata is mentioned in 37.5% of resources and operationalized in 31.2%. URLs and Internet domains support source tracing, while country-code domains (e.g., `.uk`, `.ca`) offer geographic clues without directly identifying content origin or contributor location. These gaps motivate the user-centric documentation checklist (§7).

7 Discussion: Toward User-Centric Benchmark Design

Our three audits show that evaluation resources encode user-relevant assumptions through content, metadata, documentation, and scoring procedures. Linguistic conventions affect which answers actually receive credit, geographic references reveal uneven country visibility, and documentation specifies data provenance more clearly than intended evaluation users. These findings connect resource design to benchmark validity: interpreting scores requires understanding whose conventions, contexts, and judgments define successful performance and where evaluation claims reliably apply.

A user-centric approach makes these assumptions *explicit* and *inspectable*. Benchmark creators should explicitly document whose data and judgments shape a resource, which populations and settings its evaluation claims represent, and how scoring procedures handle valid linguistic and locale variations. Table 6 translates these requirements into *six practical documentation fields*, with an extended reporting template supporting implementation in Appendix D.

Implications for benchmark design and use. For creators, the checklist supports clearly specifying intended users, documenting scoring conventions, and identifying validation needs. For benchmark users, it helps assess whether reported scores apply to specific populations, languages, locales, and deployment settings. Making these assumptions inspectable also enables researchers to critically examine unsupported generalizations and assess the scope of evaluation claims.

Auditing evaluation begins with the resource. Benchmark content, metadata, reference answers, and scoring procedures shape evaluation before models are evaluated. Making their assumptions explicit strengthens score interpretation and clarifies the scope of claims across user contexts.

8 Conclusion

We introduced a user-centric framework for auditing evaluation resources and applied it to 11 text benchmarks and five image–text resources. Our findings reveal American-centered linguistic conventions, geographic visibility associated with national GDP share, and gaps between documented data sources and intended users. Paired scoring experiments show that meaning-preserving answer variants can receive different scores. These findings highlight how assumptions shape evaluation and why making them explicit is essential for interpreting benchmark scores across user contexts.

Limitations

Linguistic conventions and scoring. Our linguistic audit examines explicit spelling, date, time, and measurement conventions, providing a reproducible account of observable locale assumptions. Paired experiments using the official DROP scorer and reproduced TruthfulQA generation evaluator quantify how meaning-preserving answer variants affect evaluation outcomes. These analyses characterize linguistic representation and scoring sensitivity within the audited resources.

Geographic visibility and socioeconomic structure. We measure geographic visibility through explicit country references in captions and associated textual metadata, using deterministic resolution to support consistent cross-resource comparisons. Country shares reflect resolved textual references, while socioeconomic correlations characterize associations between geographic visibility and national economic indicators.

Public documentation and auditability. Our documentation audit examines publicly available benchmark papers, dataset cards, and repositories. The rubric assesses whether user-relevant assumptions are explicitly documented and operationalized through inspectable evidence. Its scores therefore characterize the transparency and auditability of released documentation.

Empirical scope and extensibility. Our study examines 11 text benchmarks and five image–text resources, covering selected linguistic, geographic, and socioeconomic dimensions. The framework provides a foundation for broader audits spanning multilingual evaluation, accessibility, diverse user contexts, and richer annotation metadata.

Ethics Statement

Data use and privacy. This study audits publicly available evaluation resources, documentation, textual metadata, and socioeconomic indicators. Analyses use research materials and aggregate measurements, respecting dataset licenses and usage conditions. Geographic visibility is measured through captions and associated metadata, with country assignments representing textual references rather than image provenance or contributor locations.

Responsible interpretation. Our findings characterize linguistic and locale conventions, geographic visibility, socioeconomic associations, scoring sensitivity, and documentation practices. These measurements describe resource-level properties and evaluation procedures, supporting interpretation of benchmark assumptions without inferring individual identities or demographic characteristics. Socioeconomic relationships are reported as descriptive associations.

User-centric accountability. Our framework encourages documentation of data provenance, intended users, annotator practices, and scoring conventions. These disclosures help benchmark creators define the scope of evaluation claims and enable researchers to assess their applicability across linguistic, geographic, and user contexts.

Open science and reproducibility. We publicly release code and research artifacts, including convention detectors, country-resolution rules, aggregate results, regression analyses, and documen-

tation audit materials.[2] Our code is licensed under Apache 2.0, while original research artifacts are released under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).[3] These materials support reproducibility and future research efforts while respecting the applicable licenses and distribution restrictions of the original resources.

References

Abhipsa Basu, Yugam Bahl, Kirti Bhagat, Preethi Seshadri, R. Venkatesh Babu, and Danish Pruthi. 2026. Where do images come from? analyzing captions to geographically profile datasets. *Preprint*, arXiv:2602.09775.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. COYO-700M: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset.

Nikhil Chandak, Shashwat Goel, Ameya Prabhu, Moritz Hardt, and Jonas Geiping. 2025. Answer matching outperforms multiple choice for language model evaluation. *Preprint*, arXiv:2507.02856.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3557–3567.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

David Crystal. 2003. *English as a Global Language*, 2 edition. Cambridge University Press, Cambridge, UK. Second edition with updated discussions on the global spread of English and its sociolinguistic implications.

Karan Desai, Gaurav Kaul, Zubin Trivadi Aysola, and Justin Johnson. 2021. Redcaps: Web-curated image-text data created by the people, for the people. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does object recognition work for everyone? *Preprint*, arXiv:1906.02659.

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

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: An NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, and 15 others. 2023. Datacomp: In search of the next generation of multimodal datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach,

[2] Data, code, and research artifacts are available at tafseer-nayeem.github.io/WhoseBench.

[3] https://creativecommons.org/licenses/by-nc-sa/4.0/

Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

GeoNames. 2026. GeoNames Geographical Database. https://www.geonames.org/. Accessed April 25, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 22 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Yu Lu Liu, Su Lin Blodgett, Jackie Cheung, Q. Vera Liao, Alexandra Olteanu, and Ziang Xiao. 2024. ECBD: Evidence-centered benchmark design for NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16349–16365, Bangkok, Thailand. Association for Computational Linguistics.

John Miller. 2022. Validity challenges in machine learning benchmarks. Technical Report UCB/EECS-2022-180, Department of Electrical Engineering and Computer Sciences, University of California, Berkeley. Technical Report No. UCB/EECS-2022-180.

Seyed Mahed Mousavi, Edoardo Cecchinato, Lucia Horníková, and Giuseppe Riccardi. 2026. Garbage in, reasoning out? why benchmark scores are unreliable and what to do about it. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 1747–1759, Rabat, Morocco. Association for Computational Linguistics.

Mir Tafseer Nayeem and Davood Rafiei. 2024. KidLM: Advancing language models for children – early insights and future directions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4813–4836, Miami, Florida, USA. Association for Computational Linguistics.

Mir Tafseer Nayeem and Davood Rafiei. 2025. Opinorag: Towards generating user-centric opinion highlights from large-scale online reviews. *Preprint*, arXiv:2509.00285.

Mir Tafseer Nayeem and Davood Rafiei. 2026. How does "english (us)" become the default? triangulating structural bias towards american english across the llm pipeline. *Preprint*, arXiv:2604.04204.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336.

Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *Preprint*, arXiv:1711.08536.

Vered Shwartz. 2022. Good night at 4 pm?! time expressions in different cultures. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2842–2853, Dublin, Ireland. Association for Computational Linguistics.

Anna Sokol, Elizabeth Daly, Michael Hind, David Piorkowski, Xiangliang Zhang, Nuno Moniz, and Nitesh Chawla. 2025. Benchmarkcards: Standardized documentation for large language model benchmarks. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2443–2449, New York, NY, USA. Association for Computing Machinery.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 33 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Leon Tikly. 2016. Language-in-education policy in low-income, postcolonial contexts: Towards a social justice approach. *Comparative Education*, 52(3):408–425.

Unicode Consortium. 2026. Unicode Common Locale Data Repository (CLDR). https://cldr.unicode.org/. Accessed April 25, 2026.

Jiayin Wang, Fengran Mo, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024a. A user-centric multi-intent benchmark for evaluating large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3588–3612, Miami, Florida, USA. Association for Computational Linguistics.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024b. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024c. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

World Bank. 2026. World development indicators. https://data.worldbank.org/indicator. Accessed April 25, 2026.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Supplementary Material: Appendices

Appendix Contents

A Benchmark Convention Audit
- A.1 Benchmark Selection and Evaluation Text
- A.2 Convention Inventory and Detection Rules
- A.3 Item Labels and Aggregation Metrics
- A.4 Convention-Level Distributions
- A.5 Coverage and Spelling-Excluded Robustness
- A.6 Scorer Case Study: Construction and Method
- A.7 Scorer Case Study: Detailed Results

B Geographic Visibility Audit
- B.1 Image–Text Resources
- B.2 Country Lexicon Construction
- B.3 Matching and Country Assignment
- B.4 Geographic Concentration
- B.5 GDP and Country Representation
- B.6 Socioeconomic Comparisons and Controls
- B.7 Language-Composition Stress Test on WIT

C Documentation Audit

D User-Centric Documentation Template

E AI Use Statement

A Benchmark Convention Audit

This appendix documents the RQ1 convention audit, covering benchmark selection, detection rules, aggregation metrics, spelling-excluded robustness, and the paired scorer study (section 4).

A.1 Benchmark Selection and Evaluation Text

We audit 11 text benchmarks: MMLU, GSM8K, HellaSwag, DROP, WinoGrande, ARC-Challenge, BoolQ, MMLU-Pro, TruthfulQA, MedQA, and SVAMP. These resources span professional knowledge, mathematical reasoning, commonsense reasoning, reading comprehension, science, truthfulness, medical question answering, and robustness-oriented mathematical reasoning. Table 7 summarizes their tasks and evaluation formats.

Each benchmark item is the unit of analysis. We concatenate questions, passages, instructions, answer options, and gold answers to capture the linguistic conventions presented to readers and used in evaluation. The audit characterizes these resource texts independently of model-generated responses.

A.2 Convention Inventory and Detection Rules

We examine four classes of explicit linguistic and locale conventions:

1. **Spelling.** Orthographic alternatives, including *color*/*colour* and *center*/*centre*.
2. **Date formats.** Numeric and written date conventions, including `04/21/2025`/`21/04/2025` and *April 21, 2025*/*21 April 2025*.
3. **Time formats.** Twelve-hour and twenty-four-hour expressions, such as *6:30 pm*/`18:30`.
4. **Measurement units.** Contrasts involving U.S.-customary and metric conventions, including *miles*/*kilometres*, *pounds*/*kilograms*, and *Fahrenheit*/*Celsius*.

Spelling detection uses the curated orthographic inventory of Nayeem and Rafiei (2026). Pattern-based detectors identify date, time, and measurement conventions in benchmark text, drawing on temporal distinctions discussed by Shwartz (2022).

Detection targets explicit surface forms. Spelling variants require lexical matches, preventing partial matches inside unrelated words. Date and time expressions are identified by structural patterns; abbreviated units additionally require quantitative context. These rules provide consistent, inspectable evidence across benchmarks.

The American-centered category comprises the audited American English spellings, month–day dates, twelve-hour times, and U.S.-customary measurement signals. The comparison category contains the corresponding other user-valid forms. The labels describe the conventions, independently of the identity or nationality of item contributors. Table 8 illustrates these convention contrasts.

Benchmark	Task	Category	Short description
MMLU (Hendrycks et al., 2021)	Multiple-choice QA over 57 subjects	Broad knowledge and reasoning	Academic and professional exam questions spanning humanities, STEM, and professional domains.
GSM8K (Cobbe et al., 2021)	Grade-school math word problems	Mathematical reasoning	Linguistically diverse elementary math problems with free-form reasoning and numeric answers.
HellaSwag (Zellers et al., 2019)	Sentence / ending completion	Commonsense reasoning	Adversarially filtered everyday-event completion benchmark with multiple-choice endings.
DROP (Dua et al., 2019)	Reading comprehension with arithmetic	Reading + discrete reasoning	Paragraph-based QA requiring counting, comparison, addition, subtraction, and date/number tracking.
WinoGrande (Sakaguchi et al., 2021)	Pronoun / referent resolution	Commonsense coreference	Large-scale adversarial Winograd-style benchmark for commonsense pronoun resolution.
ARC-Challenge (Clark et al., 2018)	Science multiple-choice QA	Science reasoning	Difficult subset of grade-school science questions from the AI2 Reasoning Challenge.
BoolQ (Clark et al., 2019)	Yes/no passage QA	Factual QA	Naturally occurring yes/no questions paired with short passages, originally sourced from search queries.
MMLU-Pro (Wang et al., 2024c)	Harder multiple-choice QA	Broad knowledge + reasoning	Harder successor to MMLU with more reasoning-focused items and expanded answer-option sets.
TruthfulQA (Lin et al., 2022)	Open-ended and multiple-choice QA	Truthfulness / factuality	Questions designed to elicit common misconceptions and imitative falsehoods.
MedQA (Jin et al., 2021)	Medical exam multiple-choice QA	Domain-specific professional QA	Medical board-style exam questions requiring specialized clinical knowledge and decision-making.
SVAMP (Patel et al., 2021)	Perturbed math word problems	Robustness in math reasoning	Challenge set built by applying controlled variations to simple arithmetic word problems.

Table 7: Overview of the benchmarks used in our evaluation, summarizing each benchmark's task formulation, capability category, and intended reasoning or knowledge focus.

Convention	American-centered example	Other user-valid example	Signal type
Spelling	*color*, *center*	*colour*, *centre*	Orthographic
Numeric date	`04/21/2025`	`21/04/2025`	Date order
Written date	*April 21, 2025*	*21 April 2025*	Date order
Time notation	*7:00 pm*	`19:00`	Time format
Measurement	*5 miles*	*8.05 kilometres*	Unit convention
Temperature	*68 degrees Fahrenheit*	*20 degrees Celsius*	Unit convention

Table 8: Illustrative linguistic and locale conventions used in the text benchmark audit. The examples show alternative valid forms; the detectors measure their occurrence in benchmark text rather than assigning correctness or user identity.

A.3 Item Labels and Aggregation Metrics

For each item, we retain the detected convention signals and assign one of four labels:

$$\texttt{american_only},\quad \texttt{other_only},\quad \texttt{mixed},\quad \texttt{no_match}. \tag{1}$$

Items receive `american_only` when they contain exclusively American-centered signals and `other_only` when they contain exclusively comparison signals. Items containing both receive `mixed`; those with neither receive `no_match`.

We summarize convention prevalence using *coverage*, the proportion of items containing at least one audited signal:

$$\text{Coverage}_b = \frac{|\{i \in b : \ell(i) \neq \texttt{no_match}\}|}{N_b}, \tag{2}$$

where N_b is the number of items in benchmark b, and $\ell(i)$ is the assigned item label.

Coverage measures how frequently audited conventions appear in benchmark items. Reporting it alongside directional signal shares distinguishes the prevalence of identifiable conventions from the balance of forms among those matches.

Let A_b and O_b denote the counts of American-centered and other user-valid signal matches in benchmark b, respectively. We define:

$$\text{AmE Share}_b = \frac{A_b}{A_b + O_b}. \tag{3}$$

Mixed items may contribute to both A_b and O_b. Coverage is therefore computed over items, whereas American-centered share is computed over matched signals.

For a signed representation of the same distribution, we use:

$$\text{Skew}_b = 2\,\text{AmE Share}_b - 1. \tag{4}$$

Skew ranges from -1 to 1: positive values indicate an American-centered majority, negative values indicate a majority of comparison signals, and zero denotes equal shares.

The comparison category includes several spelling traditions, date and time formats, and measurement conventions. The aggregate share measures the balance of this defined inventory rather than identifying a single alternative English variety.

A.4 Convention-Level Distributions

Main-paper Figure 1 presents aggregate convention signals across benchmarks. Figure 4 separates spelling, dates, time, and measurement units, exposing the convention-specific patterns within each resource. Spelling contributes prominently to the American-centered profiles, with varying influence across benchmarks. Mathematics and reading-comprehension resources also contain date, time, and measurement signals, whereas some knowledge and commonsense resources are dominated by orthographic differences.

The category-level view separates each benchmark's aggregate direction from the convention classes contributing to it.

A.5 Coverage and Spelling-Excluded Robustness

We report benchmark-level coverage and American-centered share separately to distinguish signal prevalence from directional concentration. Coverage spans 5.97% in WinoGrande and 7.59% in TruthfulQA to 65.85% in MedQA and 78.08% in DROP, demonstrating substantial differences in the proportion of items containing audited signals.

We repeat the directional analysis after removing all spelling matches. For benchmark b, let $A_b^{\neg S}$ and $O_b^{\neg S}$ denote the remaining American-centered and other convention matches. We compute:

$$\text{AmE Share}_b^{\neg S} = \frac{A_b^{\neg S}}{A_b^{\neg S} + O_b^{\neg S}}. \tag{5}$$

Table 2 shows an American-centered majority after removing spelling in eight of 11 benchmarks. TruthfulQA (90.5%), GSM8K (83.3%), SVAMP

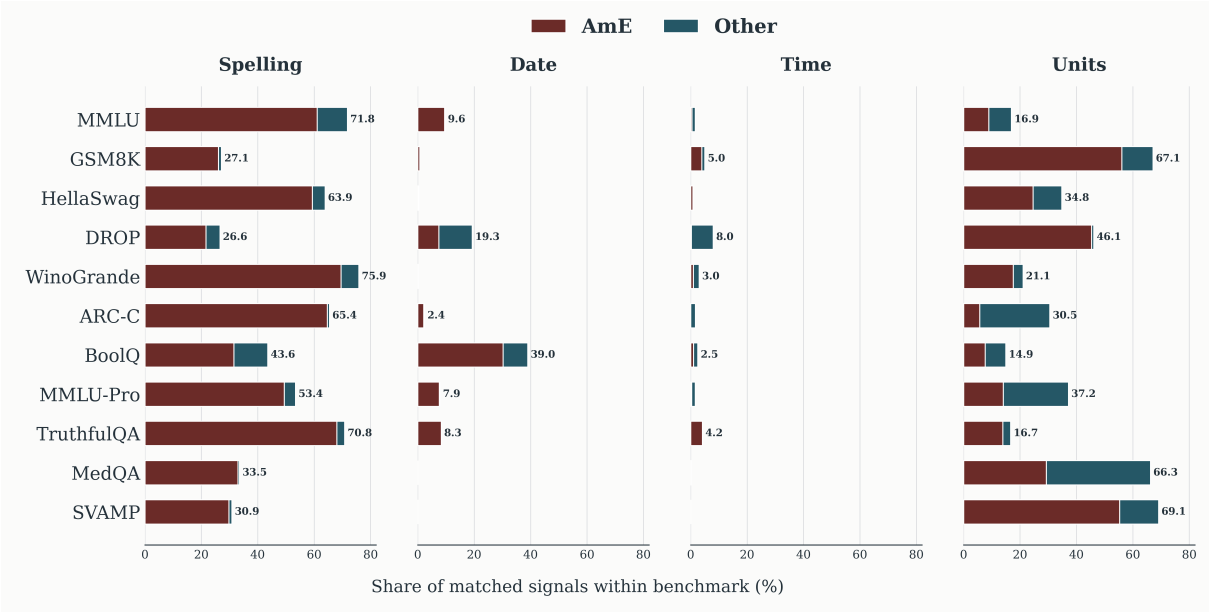

Figure 4: Benchmark-level distribution of linguistic and locale convention signals across 11 text benchmarks. Each panel shows, for a convention type, the proportions of American-centered and other user-valid matches within each benchmark. American-centered forms predominate in many benchmarks, with substantial variation in the direction of matches across spelling, dates, time, and measurement units. Date and time signals are generally sparser, revealing distinct convention profiles across resources.

(80.0%), and DROP (72.3%) retain high shares. ARC-Challenge (22.5%), MedQA (44.3%), and MMLU-Pro (47.3%) instead have majorities of other convention signals.

These results identify benchmarks whose aggregate profiles depend strongly on spelling and establish that the American-centered pattern also appears in nonspelling conventions.

A.6 Scorer Case Study: Construction and Method

We extend the convention audit from benchmark content to scoring behavior by testing how answer-text evaluators respond to meaning-preserving changes in linguistic or locale conventions.

Benchmark scoring procedures. We study DROP (Dua et al., 2019) using its official exact-match (EM) and token-level F1 implementation, including normalization rules, and TruthfulQA (Lin et al., 2022) using the released generation evaluator's BLEU/ROUGE ranking of correct against incorrect reference answers.

These two evaluation interfaces enable controlled answer-text comparisons. Multiple-choice labels and extracted numerical answers use different scoring representations and are evaluated separately in their respective benchmark protocols.

Candidate construction and screening. Each pair consists of an original reference-consistent answer and a meaning-preserving variant in regional spelling, written-date order, or time notation. We score both against identical gold annotations, isolating the effect of answer expression.

For DROP, we screen 23 answer-span candidates. Eleven originals do not receive full credit because they are incomplete gold answers. The resulting analysis contains 12 pairs from 11 questions: six spelling variants and six written-date variants.

For TruthfulQA, we screen 37 candidates from 27 questions. Three substitutions alter the brand name *Flavor-Aid*, and one changes the official name *World Health Organization*. Removing these four semantic changes yields 33 pairs from 25 questions: 28 spelling, three time-format, and two written-date variants.

These screening decisions preserve referenced entities and facts, ensuring that the paired test isolates differences in convention.

DROP scoring. We apply DROP's official EM and F1 implementation to each original and variant. Originals are selected for full credit, yielding a common baseline. EM checks normalized answer agreement, while token-level F1 assigns partial credit for overlap.

TruthfulQA scoring. We reproduce the released generation evaluator's BLEU/ROUGE comparison between correct and incorrect references. Let R_q^+ and R_q^- denote the sets of correct and incorrect answers for question q, and let $m(y, r)$ be the similarity between candidate answer y and reference r under metric m. The correctness indicator is:

$$A_m(y, q) = \mathbf{1}\{ \max_{r \in R_q^+} m(y, r) > \max_{r \in R_q^-} m(y, r)\}. \tag{6}$$

The highest correct-reference similarity must strictly exceed the incorrect-reference similarity for a positive correctness decision. We preserve the released evaluator's correct reference "I have no comment." when computing the ranking.

BLEU follows the released evaluator's sacreBLEU configuration and ROUGE uses ROUGE-score F1. For each original and variant, we compute the binary correctness indicator and the maximum similarity to a correct reference. The case study evaluates these released BLEU/ROUGE procedures; BLEURT and GPT-judge are separate evaluation interfaces.

Paired differences and uncertainty. For every scoring metric, we calculate the variant-minus-original difference for each answer pair and report the mean paired change.

We estimate 95% percentile confidence intervals with 10,000 question-clustered bootstrap replicates (seed 2026). Resampling questions keeps all variants from the same source question in one cluster, accounting for within-question dependence.

We report DROP EM and F1 separately and distinguish TruthfulQA binary correctness from continuous reference similarity. Benchmark-specific reporting preserves the different meanings and scales of the evaluation metrics.

A.7 Scorer Case Study: Detailed Results

Table 9 presents the complete paired results, including pair and question counts, original and variant scores, and question-clustered confidence intervals.

DROP: exact match and partial credit. On 12 eligible pairs, official EM falls from 100.0% to 33.3%, a paired change of -66.7 percentage points (95% CI $[-92.3, -36.4]$). Official F1 falls from 100.0% to 77.2%, a change of -22.8 points (95% CI $[-41.6, -5.2]$).

Transformation type determines the amount of partial credit. The six spelling pairs lose an average of 45.7 F1 points (95% CI $[-72.2, -23.0]$); written-date variants retain full F1 even when EM decreases. Thus, token overlap recognizes valid alternatives that normalized exact match does not.

TruthfulQA: similarity and correctness. Across 33 answer pairs, BLEU and ROUGE-1 correctness each remain at 97.0%, and ROUGE-Lsum correctness remains at 100.0%. ROUGE-2 correctness decreases from 100.0% to 93.9%, a 6.1-point change (95% CI $[-16.1, 0.0]$).

Continuous reference similarity changes more substantially. Maximum correct-reference BLEU similarity falls from 97.0 to 69.2 (change -27.7; 95% CI $[-34.0, -21.9]$), while maximum ROUGE-1 similarity falls from 100.0 to 88.9 (change -11.1; 95% CI $[-13.5, -9.0]$). Both are measured on the reported 0–100 scale.

TruthfulQA's reference ranking can retain a correctness judgment as similarity to the closest correct reference decreases. This contrasts with the EM and partial-credit responses observed in DROP.

Supplementary exact-reference diagnostic. As an additional diagnostic, we test normalized exact-reference matching on all 33 TruthfulQA variants. This strict matching rule rejects every transformed answer, illustrating sensitivity to answer expression. The diagnostic is separate from TruthfulQA's released generation metrics and is not reported as official TruthfulQA correctness.

Implications for evaluation normalization. The paired results expose four distinct responses to meaning-preserving variation: loss of exact-match credit, partial token-overlap credit, reduced reference similarity, and unchanged reference-ranking correctness.

These findings motivate including meaning-preserving answer alternatives in scorer audits and documenting accepted spelling, date, time, and measurement conventions alongside evaluation rules. Such checks make recognition of equivalent responses measurable while retaining the underlying correctness criterion.

Benchmark	Scoring rule	Pairs	Questions	Original	Variant	Change [95% CI]
DROP	Official EM	12	11	100.0%	33.3%	−66.7 [−92.3, −36.4]
DROP	Official F1	12	11	100.0%	77.2%	−22.8 [−41.6, −5.2]
TruthfulQA	BLEU correctness	33	25	97.0%	97.0%	0.0 [0.0, 0.0]
TruthfulQA	ROUGE-1 correctness	33	25	97.0%	97.0%	0.0 [0.0, 0.0]
TruthfulQA	ROUGE-2 correctness	33	25	100.0%	93.9%	−6.1 [−16.1, 0.0]
TruthfulQA	ROUGE-Lsum correctness	33	25	100.0%	100.0%	0.0 [0.0, 0.0]
TruthfulQA	BLEU correct-reference similarity	33	25	97.0	69.2	−27.7 [−34.0, −21.9]
TruthfulQA	ROUGE-1 correct-reference similarity	33	25	100.0	88.9	−11.1 [−13.5, −9.0]

Table 9: Detailed paired scorer sensitivity to meaning-preserving answer variants. DROP uses its official EM/F1 scorer; TruthfulQA reproduces the released BLEU/ROUGE correct-versus-incorrect reference-ranking procedure. EM, F1, and correctness changes are percentage points; reference similarities are reported on a 0–100 scale. Confidence intervals are calculated using 10,000 question-clustered bootstrap replicates. Benchmark results are not pooled.

B Geographic Visibility Audit

This appendix documents the RQ2 audit of country visibility in five image–text resources, covering the data sources, geographic lexicon, country resolver, concentration analyses, socioeconomic comparisons, and WIT language-composition stress test (section 5). Geographic visibility throughout refers to explicit references in captions and associated textual metadata.

B.1 Image–Text Resources

We analyze DataComp1B, WIT English, RedCaps, CC12M, and COYO-700M, representing web-scale image–text data, Wikipedia-derived metadata, Reddit captions, web alt-text, and CommonCrawl-derived pairs. Table 10 documents their characteristics and roles in the audit.

We analyze the textual component of each resource. A record is *location-specified* when captions or associated metadata supply sufficient evidence for a country assignment; otherwise it is *location-underspecified*.

Country visibility is calculated among location-specified records. The measure captures places referenced by captions and metadata, including contexts associated with described subjects, events, and entities. It measures geographic references in text rather than the geographic origin of photographs or contributors.

B.2 Country Lexicon Construction

Our geographic lexicon contains 3,413 entries: 3,036 country-name entries, 83 demonyms, and 294 city entries. Each record stores `country_alpha2`, `country_alpha3`, `canonical_name`, `alias`, `alias_norm`, `source`, `match_type`, and `confidence_tier`.

The lexicon combines four sources:

- **Country names.** ISO-style records from `pycountry`[4], including canonical and official names and their diacritic-folded forms.
- **Localized names.** Territory names from Babel/CLDR for English, French, German, Spanish, Portuguese, Japanese, Hindi, Greek, Simplified Chinese, and Traditional Chinese (Unicode Consortium, 2026).
- **Aliases and demonyms.** Manually curated country abbreviations, alternative names, and adjectival references.
- **Cities and capitals.** National capitals from GeoNames through `geonamescache`[5], supplemented with selected high-signal cities (GeoNames, 2026).

Country aliases include `USA`, `UK`, `UAE`, and `South Korea`; demonyms include `French`, `Japanese`, `Indian`, and `Brazilian`. The city layer covers Paris, London, Tokyo, New York, Mumbai, Sydney, and other selected locations.

We filter ambiguous city aliases, excluding digit-containing names, very short ASCII alternatives, common words, capital aliases shared across countries, and place names frequently used outside geographic contexts.

All aliases undergo Unicode NFKC normalization, lowercasing, punctuation and symbol normalization, and whitespace collapsing. Country names with diacritics retain original and ASCII-folded forms where available.

[4] https://github.com/pycountry/pycountry
[5] https://github.com/yaph/geonamescache

Benchmark	Source / formulation	Role in this study	Short description
DataComp1B (Gadre et al., 2023)	Web-scale image–text data curation benchmark.	Core web training-corpus benchmark.	Broad, noisy web image–text source used to test whether country mentions in caption metadata reflect the geographic skew of large-scale multimodal training data.
WIT English (Srinivasan et al., 2021)	Wikipedia image–text dataset with captions and structured page context.	Curated encyclopedia-style control.	English view of WIT, using caption and contextual text fields from Wikipedia. It provides a more structured source domain than web-crawl alt-text corpora.
RedCaps (Desai et al., 2021)	Reddit-sourced image–text pairs from curated subreddits.	Social-web comparison benchmark.	User-generated social-media caption source. It tests whether geographic concentration persists outside Wikipedia and conventional web alt-text datasets.
CC12M (Changpinyo et al., 2021)	Web-collected image–alt-text pairs extending Conceptual Captions.	Classic alt-text caption benchmark.	Large web caption corpus designed to broaden Conceptual Captions toward long-tail visual concepts. It gives a canonical vision–language pretraining comparison point.
COYO-700M (Byeon et al., 2022)	CommonCrawl-derived image–text pairs with additional metadata and filtering scores.	Web-scale expansion benchmark.	Open web-scale image–text corpus included to test whether the observed geographic patterns remain consistent beyond the smaller benchmark and control datasets.

Table 10: Overview of the five image–text resources included in the geographic visibility audit. The table summarizes their data sources, characteristics, and roles in the analysis. Geographic visibility is measured using explicit country references resolved from captions and associated textual metadata. Country assignments reflect textual references rather than image geolocation or the geographic origin of the content.

Each lexicon entry is typed as a country name, demonym, or city; that type determines its evidence weight in country assignment.

B.3 Matching and Country Assignment

The resolver detects geographic mentions in captions and metadata and aggregates evidence for each candidate country.

ASCII aliases are matched as complete token sequences; non-ASCII aliases are matched in normalized Unicode text. Repeated matches of the same alias, country, and type within one record are deduplicated.

For each candidate country c, we calculate:

$$s(c) = 4n_{\text{country}}(c) + 3n_{\text{demonym}}(c) + 2n_{\text{city}}(c), \tag{7}$$

where the three counts represent deduplicated country-name, demonym, and city matches.

The weighting scheme prioritizes explicit country names over demonyms and city references. Cities provide useful geographic evidence but require more contextual discrimination.

Each record receives one of four outcomes:

- **Single-country:** All matched geographic evidence identifies one country.
- **Weighted-majority:** Multiple countries are mentioned, but one receives a uniquely highest weighted score.
- **Ambiguous:** Multiple countries receive competing evidence without a unique highest score.
- **Unresolved:** No lexicon entry is matched.

Country frequencies include single-country and uniquely highest weighted-majority assignments. Ambiguous and unresolved records are excluded from the denominator, and ISO alpha-2 codes are mapped to canonical country names for reporting.

Country visibility describes the distribution of resolved textual assignments. Resolved coverage separately measures the proportion of records whose text supplies enough evidence for assignment.

B.4 Geographic Concentration

Main-paper Figure 2 reports the five leading countries per resource; Figure 5 extends these distributions to the top ten.

The United States leads in DataComp1B (13.55%), WIT English (14.59%), RedCaps (21.18%), and COYO-700M (16.30%) among resolved assignments.

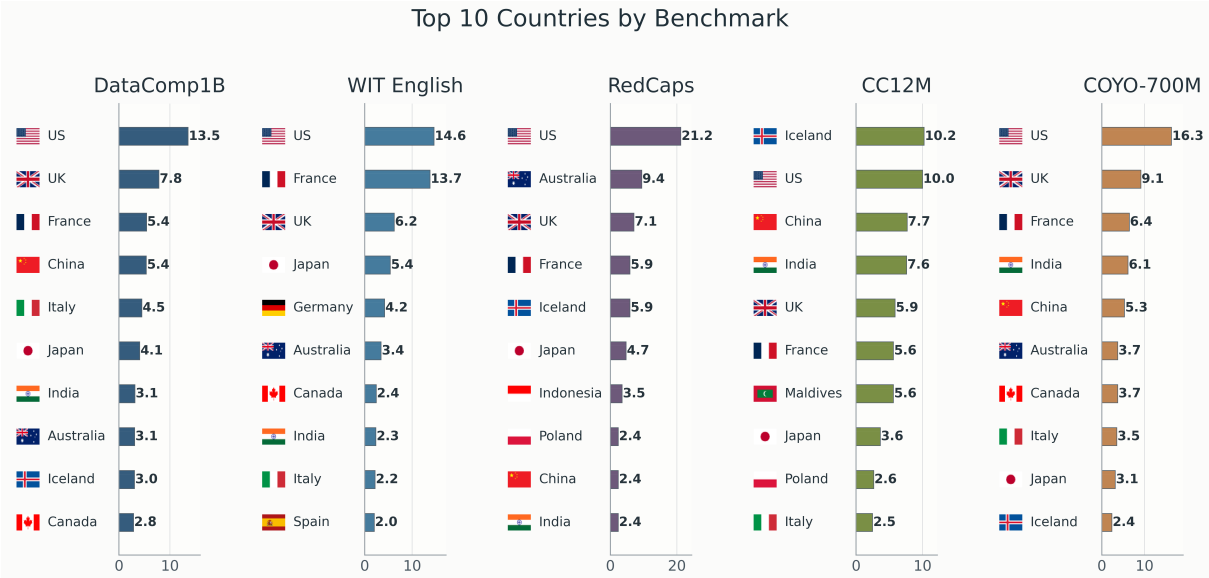

Figure 5: **The high-frequency tail remains geographically concentrated.** Extending Figure 2 to the top ten countries shows that resolved caption geography often centers on recurring countries in North America, Europe, and large Asian economies. CC12M is a benchmark-specific exception in which Iceland is the top resolved country, suggesting that source-pipeline artifacts or repeated content clusters may shape some country-frequency patterns. Rankings are benchmark-specific and do not require cross-benchmark overlap.

CC12M exhibits a different ranking: Iceland (10.24%), the United States (9.99%), China (7.68%), India (7.60%), and the United Kingdom (5.86%). This profile highlights variation across collection sources; repeated domains, alt-text conventions, or content clusters are potential contributors to the observed distribution.

Across the top-ten results, frequently referenced countries include the United States, United Kingdom, France, Australia, Canada, Japan, India, and China. These countries span North America, Europe, and major Asian economies.

Table 11 complements the leading-country view by listing the lowest-frequency countries with at least two resolved assignments in each resource.

The low-frequency lists span Africa, Asia, Europe, the Americas, the Caribbean, and the Pacific. Rankings are computed independently within each resource and reveal a geographically diverse tail of sparsely referenced countries. The contrast with the recurring leading countries illustrates concentration in the textual geography of these collections.

B.5 GDP and Country Representation

We link resolved country assignments to World Bank World Development Indicators through ISO identifiers and quantify their relationship with socioeconomic structure (World Bank, 2026).

For benchmark b and country c, let $d_{c,b}$ denote the country's share of resolved assignments and $g_{c,b}$ its share of GDP among the countries retained for that benchmark. Our primary regression is:

$$\log_{10} d_{c,b} = \alpha_b + \beta_b \log_{10} g_{c,b} + \epsilon_{c,b}. \quad (8)$$

We report log-space Pearson correlations, Spearman rank correlations, count-weighted correlations, and ordinary least-squares R^2. Count weighting uses the number of resolved assignments for each country in correlations and regressions.

Socioeconomic indicators. We retrieve GDP in current U.S. dollars (`NY.GDP.MKTP.CD`), GDP per capita (`NY.GDP.PCAP.CD`), population (`SP.POP.TOTL`), Internet usage (`IT.NET.USER.ZS`), and international tourism arrivals (`ST.INT.ARVL`).

GDP, GDP-per-capita, and population observations are predominantly from 2024; available tourism observations often refer to 2019 or 2020.

Country shares are normalized over aligned benchmark-specific sets with available required indicators. Regressions with Internet usage or tourism use their corresponding complete-case country sets.

Our primary specification retains countries with populations of at least one million. This threshold limits disproportionate effects from very small economies with unusually high tourism or web visibility relative to GDP and retains all five resources.

Rank	DataComp1B	WIT English	RedCaps	CC12M	COYO-700M
1	Belarus (2)	Angola (2)	Anguilla (2)	Afghanistan (2)	Argentina (2)
2	Belize (2)	Bahrain (2)	Canada (2)	Bahamas (2)	Bangladesh (2)
3	Bolivia (2)	Belize (2)	China (2)	Belize (2)	Bulgaria (2)
4	Congo (2)	Benin (2)	Germany (2)	Brunei Darussalam (2)	Ethiopia (2)
5	Dominica (2)	Cameroon (2)	India (2)	Denmark (2)	Holy See (2)
6	Guam (2)	Dominica (2)	Italy (2)	Falkland Islands (2)	Jamaica (2)
7	Lebanon (2)	Guyana (2)	Maldives (2)	Fiji (2)	Kuwait (2)
8	Malta (2)	Jamaica (2)	Poland (2)	Guernsey (2)	Moldova (2)
9	Martinique (2)	Jordan (2)	Indonesia (3)	Iran (2)	Monaco (2)
10	Monaco (2)	Lao People's Democratic Republic (2)	Japan (4)	Jamaica (2)	Nicaragua (2)

Table 11: **Benchmark-wise low-frequency countries with at least two samples.** Each column lists the ten lowest-share countries within a benchmark, ranked independently over resolved country assignments. Countries are included only when they have at least two samples in that benchmark; counts in parentheses denote those caption-derived assignments. The resulting lists include small island territories, Caribbean and Pacific countries, and countries from Africa, Asia, Europe, and the Americas. This pattern suggests a broad long tail of sparsely represented geographies rather than a single region-specific effect.

Benchmark	Countries	Pearson r	Spearman ρ	Weighted r	OLS R^2
DataComp1B	118	0.816	0.786	0.907	0.667
WIT English	146	0.851	0.823	0.874	0.724
RedCaps	30	0.647	0.735	0.809	0.419
CC12M	74	0.770	0.772	0.833	0.593
COYO-700M	109	0.817	0.775	0.897	0.667
Mean	–	**0.780**	**0.778**	**0.864**	**0.614**

Table 12: GDP-share association with country visibility across five image–text resources. Countries with populations below one million are excluded. Pearson correlations use log GDP share and log dataset share; weighted correlations use resolved country counts.

Table 12 shows positive GDP-share associations across all five resources. Spearman correlations range from 0.735 in RedCaps to 0.823 in WIT English. Mean OLS fit is $R^2 = 0.614$, compared with count-weighted $R^2 = 0.748$ reported in Table 13.

Countries with larger shares of economic activity generally receive greater visibility among the resolved textual country assignments.

B.6 Socioeconomic Comparisons and Controls

We compare GDP share with population share, GDP per capita, Internet usage, and international tourism arrivals to characterize the socioeconomic associations in country visibility.

The response variable is:

$$y_{c,b} = \log_{10} d_{c,b}, \tag{9}$$

where $d_{c,b}$ is the dataset share defined in section B.5.

Single-predictor regressions use the logarithm of GDP share, population share, GDP per capita, Internet usage, or tourism arrivals. GDP and population shares are normalized within the retained resource-specific country set.

We additionally fit multivariable models that combine GDP share with each control. For example, let u_c denote the percentage of individuals using the Internet in country c. The GDP-plus-Internet specification is:

$$\begin{aligned} y_{c,b} = \alpha_b &+ \beta_1 \log_{10} g_{c,b} \\ &+ \beta_2 \log_{10} u_c + \epsilon_{c,b}. \end{aligned} \tag{10}$$

The other multivariable specifications replace Internet usage with population share, GDP per capita, or tourism arrivals. Weighted regressions use resolved country counts as observation weights.

Table 13 shows that GDP share has the strongest individual fit among the examined indicators. Tourism arrivals ranks second, followed by population share, GDP per capita, and Internet usage.

Adding a control to GDP share produces modest improvements: mean OLS R^2 rises from 0.614 to 0.617–0.630 across the combined models.

Model	Mean OLS R^2	Weighted R^2
GDP share only	0.614	0.748
Population share only	0.252	0.374
GDP per capita only	0.210	0.254
Internet users (%) only	0.111	0.124
Tourism arrivals only	0.347	0.396
GDP share + population share	0.617	0.763
GDP share + GDP per capita	0.617	0.763
GDP share + Internet users (%)	0.629	0.751
GDP share + tourism arrivals	0.630	0.770

Table 13: Mean explanatory fit of socioeconomic indicators across the five resources. Results follow the population threshold and complete-case requirements of the geographic analysis.

These relationships are consistent with differences in economic scale and in the composition of web- and Wikipedia-derived collections. Source language and collection practices also influence the references available for geographic resolution.

B.7 Language-Composition Stress Test on WIT

We use WIT to examine geographic visibility under a different language composition, constructing balanced samples in French, German, Spanish, and Japanese. We apply the country-resolution procedure from section B.3 to each language setting. Country shares are computed from records whose caption or associated metadata yields a resolved country assignment.

Cross-language geographic matching. The shared resolver detects country names, aliases, demonyms, and selected city references in multilingual text. Place names such as *France*, *Paris*, *Berlin*, *Canada*, *Mexico*, *Chile*, and *Argentina* appear in similar forms across several Latin-script languages. Japanese records may also contain romanized place names, URLs, titles, entity names, and English aliases. Applying the same resolver keeps the detection inventory consistent across samples, enabling controlled comparison of resolved geographic references.

Geographic visibility across languages. Table 14 presents the language-composition comparison. The U.S. share falls from 20.0% in the corresponding English comparison to 8.5% in the balanced non-English sample, demonstrating that language composition materially changes the observed country distribution.

Economic concentration remains substantial: the ten countries with the largest GDP shares account for 53.5% of resolved assignments in the non-English sample, with Pearson $r = 0.752$ and Spearman $\rho = 0.744$. The comparison identifies two concurrent patterns: changing the language composition shifts the visible countries, while economically prominent countries continue to account for a substantial share of resolved geographic references.

Measure	English	Balanced non-English
U.S. share (%)	20.0	8.5
Top-10 GDP share (%)	–	53.5
Pearson r	–	0.752
Spearman ρ	–	0.744

Table 14: Geographic visibility in the WIT language-composition stress test. The balanced non-English sample includes French, German, Spanish, and Japanese. Country shares use resolved assignments; GDP correlations and top-ten concentration are reported for the non-English setting.

Interpretation. Prior work documents concentrated language and source composition in English web corpora such as C4 (Dodge et al., 2021). Our WIT analysis extends this concern to textual geography in image–text collections, showing that the language of captions and metadata is associated with which countries become visible.

The main WIT English audit and the language-comparison experiment have distinct analysis settings. The U.S. shares of 14.59% in the main audit and 20.0% in the separate English comparison therefore refer to different samples and are reported independently. Reporting geographic coverage alongside language composition and collection sources makes these influences inspectable and supports more precise interpretation of cross-resource country distributions.

C Documentation Audit

This appendix documents the RQ3 documentation audit, including the annotation protocol, inter-annotator agreement, and resource-level results (section 6). The central measure is whether public documentation states user-relevant assumptions and makes them inspectable.

Annotation procedure. Two computer science doctoral students, one author and one external volunteer, independently reviewed public documentation for all 16 resources, including papers, official dataset cards, repositories, and the sources listed

in Table 16. Both annotators had experience in NLP evaluation, benchmark analysis, and dataset documentation. Both voluntarily agreed to participate and consented to sharing their annotation data. Before annotation, both reviewers established a shared interpretation of the six criteria and their three-level scoring rubric:

- **0 (Absent):** The relevant information is not documented.
- **1 (Mentioned):** The documentation explicitly describes the relevant context or assumption.
- **2 (Operationalized):** The information is available through an inspectable metadata field, source partition, annotation protocol, scoring rule, or comparable mechanism.

All nonzero scores require explicit documentary evidence. Where several sources exist for a resource, we retain the highest evidence-supported score. Operationalization requires an inspectable field, protocol, rule, or other mechanism that a benchmark user can apply.

Source and geographic metadata. This criterion covers website URLs, Internet domains, top-level domains (TLDs), language tags, and country or region fields that support provenance tracing and geographic auditing. A *domain* here is an Internet website domain, such as `bbc.co.uk`; its `.uk` country-code TLD is associated with the United Kingdom. Likewise, `.ca` and `.jp` indicate country-code associations with Canada and Japan, whereas a generic TLD such as `.com` does not specify a country. Such signals can help trace sources while remaining distinct from content creation or contributor locations.

Inter-annotator agreement. Table 15 gives agreement before adjudication: exact agreement is 99.0% across criterion-level decisions, with overall Cohen's $\kappa = 0.984$.

The single disagreement concerns CC12M's released image URLs. One annotator judged these sufficient to operationalize source/geographic metadata; the other assigned a lower score. Adjudication assigned score 2 because the released URLs permit direct inspection of source provenance. The high agreement reflects the rubric's evidence-based criteria and supports consistent interpretation of documentation availability across resources.

Criterion	Exact (%)	κ	Disagr.
Source-user context	100.0	–	0
Target/eval. context	100.0	–	0
Annotator visibility	100.0	1.000	0
Locale assumptions	100.0	1.000	0
Evaluation normalization	100.0	1.000	0
Source/geographic metadata	93.8	0.881	1
Overall (micro)	**99.0**	**0.984**	**1**

Table 15: Inter-annotator agreement across 16 resources before adjudication. Exact agreement and Cohen's κ are reported for each criterion. Dashes indicate undefined κ when both annotators assign a constant score.

Resource-level documentation results. Table 16 identifies the consulted documentation, and Table 17 reports all six criterion scores for each resource.

The resource-level scores show a systematic source–target gap. All 16 resources operationalize source-user context through documented collection sources such as examinations, Wikipedia, Reddit, and web content. All mention target-user or evaluation context, but none operationalize the link between evaluation scores and intended user populations or settings.

Annotator visibility and locale assumptions are less consistently documented. Evaluation normalization is comparatively well specified, particularly for text benchmarks, although the accepted alternatives and scoring behaviors vary across evaluation procedures. These results establish that source provenance and the intended scope of evaluation claims are documented at different levels of specificity. Reporting both makes benchmark interpretation more transparent and supports the user-centric checklist in section 7.

D Example User-Centric Documentation Template

The checklist in Table 6 can be implemented through the following reporting fields. For each field, benchmark creators should document the available evidence, relevant assumptions, and information needed to interpret evaluation results.

1. **Source users.** Describe who produced the benchmark content and through which platforms, institutions, tasks, or collection processes. Indicate whether source populations are directly observed, inferred, or unknown.

Resource	Family	Primary documentation source
MMLU	Text QA	Hendrycks et al. (2021), paper.
GSM8K	Text QA	Cobbe et al. (2021), paper and dataset release.
HellaSwag	Text QA	Zellers et al. (2019), paper and supplement.
DROP	Text QA	Dua et al. (2019), paper.
WinoGrande	Text QA	Sakaguchi et al. (2021), paper and project page.
ARC-Challenge	Text QA	Clark et al. (2018), paper.
BoolQ	Text QA	Clark et al. (2019), paper.
MMLU-Pro	Text QA	Wang et al. (2024c), paper and supplement.
TruthfulQA	Text QA	Lin et al. (2022), paper.
MedQA	Text QA	Jin et al. (2021), paper.
SVAMP	Text QA	Patel et al. (2021), paper.
DataComp1B	Image–text	Gadre et al. (2023), DataComp paper and dataset card.
WIT English	Image–text	Srinivasan et al. (2021), paper and dataset repository.
RedCaps	Image–text	Desai et al. (2021), paper and dataset card.
CC12M	Image–text	Changpinyo et al. (2021), Conceptual 12M paper and official repository.
COYO-700M	Image–text	Byeon et al. (2022), dataset card and repository.

Table 16: Resources included in the documentation audit and the public documentation sources used for scoring. Each link points to the benchmark paper, official dataset card, or official repository consulted during the audit. All documentation sources were accessed in April 2026.

Resource	Family	Source	Target/eval.	Annot.	Locale	Eval.	Src./Geo.	Total
MMLU	Text QA	2	1	0	0	2	0	5/12
GSM8K	Text QA	2	1	1	0	2	0	6/12
HellaSwag	Text QA	2	1	1	0	2	0	6/12
DROP	Text QA	2	1	1	0	2	0	6/12
WinoGrande	Text QA	2	1	1	0	2	0	6/12
ARC-Challenge	Text QA	2	1	0	0	2	0	5/12
BoolQ	Text QA	2	1	1	0	2	0	6/12
MMLU-Pro	Text QA	2	1	1	0	2	0	6/12
TruthfulQA	Text QA	2	1	1	0	2	0	6/12
MedQA	Text QA	2	1	0	1	2	1	7/12
SVAMP	Text QA	2	1	0	0	2	0	5/12
DataComp1B	Image–text	2	1	0	0	1	2	6/12
WIT English	Image–text	2	1	0	2	0	2	7/12
RedCaps	Image–text	2	1	1	1	0	2	7/12
CC12M	Image–text	2	1	0	0	0	2	5/12
COYO-700M	Image–text	2	1	1	1	0	2	7/12

Table 17: Benchmark-wise documentation audit scores across 16 resources. Each criterion is scored as 0 (absent), 1 (mentioned), or 2 (operationalized), with a maximum total of 12. Source = source-user context; Target/eval. = target-user or evaluation context; Annot. = annotator visibility; Locale = locale assumptions; Eval. = evaluation normalization; Src./Geo. = source/geographic metadata.

2. **Target users and evaluation context.** Specify the users, settings, and evaluative claims the benchmark is intended to represent. Identify contexts where its applicability remains uncertain, including different linguistic, regional, or domain-specific settings.

3. **Annotators and label provenance.** Describe who produced labels, reference answers, or quality judgments, including available information about expertise, locale, recruitment, compensation, and quality control. For automatically derived labels, report their source and validation procedure.

4. **Locale conventions.** Document spelling, date, time, measurement, currency, numeric-format, language-variety, and source-language conventions in benchmark content and scoring. Distinguish conventions inherited from source

data from those introduced during benchmark construction.

5. **Evaluation normalization.** Specify how scoring procedures handle valid alternative answers, including differences in spelling, formatting, time, dates, and measurement conventions. Document which forms are accepted, normalized, converted, or rejected, and whether semantically equivalent responses receive consistent scores.

6. **Source and geographic metadata.** Report available source URLs, website domains, top-level domains, language tags, country or region fields, and related provenance information. Clarify whether these fields identify a content source, referenced location, depicted entity, or another geographic signal.

7. **Known coverage gaps.** Identify gaps in user populations, linguistic conventions, geographic visibility, task settings, or annotation provenance that affect the interpretation of benchmark results.

8. **Recommended and non-recommended uses.** State which evaluation tasks and comparisons the benchmark supports, and which broader claims require additional evidence, particularly for multilingual, cross-regional, or deployment settings.

9. **Auditability and updates.** Identify the released metadata, documentation, scoring code, and other resources needed to reproduce or extend the audit. Describe how changes to benchmark content, scoring rules, or documentation are recorded.

E AI Use Statement

Generative AI tools were used in limited assistive roles for **(1) text editing**, including grammar, spelling, phrasing, and clarity, and **(2) code assistance**, including routine implementation, data preprocessing, filtering, and debugging.

All AI-assisted text and code were reviewed and refined by the authors, and AI-assisted code was manually verified and tested before use. The authors independently formulated the research questions, developed the methodology, designed and conducted the experiments, analyzed and interpreted the results, and determined the final conclusions. The authors take full responsibility for the content of this work.