How Does "English (US)" Become the Default? Triangulating Structural Bias Towards American English Across the LLM Pipeline

Anonymous authors
Paper under double-blind review

Abstract

Large language models (LLMs) are increasingly embedded in educational, professional, and public infrastructure, yet widely used platforms expose "English (US)" as a primary English setting despite the global diversity of English. We ask: *How does "English (US)" become the default?* We study this question as *structural bias*, examining how geopolitical histories of data curation, digital dominance, and linguistic standardization intersect with the LLM development pipeline. Using British English as a controlled reference, we construct a curated resource of 1,813 matched American English (AmE)–British English (BrE) variants and introduce DiAlign, a dynamic, training-free method for estimating regional alignment from distributional evidence. We triangulate the AmE preference across *data exposure* $\rightarrow$ *representation* $\rightarrow$ *generation*, jointly examining pretraining and post-training data, tokenizer behavior and provenance, model prediction cost, and generated language across developer countries, prompt conditions, domains and sources, linguistic categories, and registers. AmE is consistently favored across all six audited pretraining corpora and 21 post-training datasets, is generally represented more compactly by tokenizers, and receives lower prediction cost. It also remains the dominant generation default under neutral English prompting; British-English prompting shifts this preference toward BrE but does not consistently eliminate the AmE default. To our knowledge, this is the first rigorous pipeline-wide study of *structural bias* across major phases of LLM development. Our findings show that contemporary LLMs privilege AmE as the *de facto* norm, raising concerns about linguistic homogenization, epistemic injustice, and inequity in global AI deployment, while providing a rigorous basis for targeted component-level intervention.

1 Introduction

Large language models (LLMs) are rapidly expanding across technical, professional, educational, legal, and public-sector applications (Lai et al., 2024; Kulal et al., 2024; Shahzad et al., 2025; Sajadieh et al., 2026). AI is now used in at least one area of government in 35 of 36 OECD countries (OECD, 2026), while national AI strategies increasingly emphasize sovereign control over compute, data, models, and infrastructure (Reuel et al., 2025; Sajadieh et al., 2026). As LLMs become part of institutional infrastructure, their defaults can operate at global scale. Language is one such default: English remains the most effective language for interacting with many LLMs (Behzad et al., 2024), yet widely used platforms such as ChatGPT and Claude expose "`English (US)`" as a primary English setting. This motivates our central question: *How does "English (US)" become the default?* If LLMs implicitly treat American English (AmE) as the default or normative form, the concern extends beyond stylistic preference to linguistic homogenization and uneven user experience worldwide (Bender et al., 2021). Answering this question requires rigorous pipeline-wide investigation that can localize the asymmetry and identify where component-level improvements should be directed.

We study this problem as *structural bias*: a systematic linguistic advantage that appears across multiple components of model development, not only in generated text. Sociolinguistic research shows that power and prestige shape which linguistic forms are legitimized or marginalized (Labov, 1972; 2006). In modern LLMs, these processes intersect with geopolitical histories of data curation, digital dominance, and linguistic standardization. We therefore triangulate the AmE preference to

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

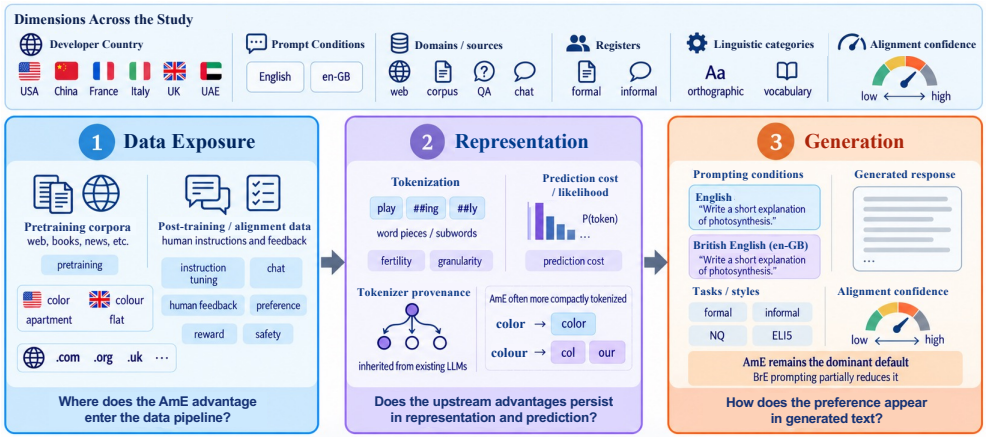


Figure 1: **Triangulating how "English (US)" becomes the default across the LLM pipeline.** We trace this asymmetry across three connected stages: **data exposure** → **representation** → **generation**, using British English as the controlled reference, Ref(BrE). The study jointly examines pretraining and post-training data, tokenizer behavior and provenance, model prediction cost, and generated language. This triangulation reveals where the American English (AmE) advantage first becomes visible, whether it persists across components, and how it manifests in model outputs. Its strength and form are examined across pipeline stages, developer countries (e.g., USA, UK, China, France, Italy, and UAE), prompt conditions (e.g., default English vs. British English), domains and sources (e.g., web, books, encyclopedic text, and mixed-domain corpora), linguistic categories (e.g., orthographic vs. vocabulary contrasts), registers (e.g., formal/encyclopedic vs. informal/conversational), and alignment confidence levels. This breadth provides a deeper, pipeline-level view of *structural bias*.

demystify the "English (US)" default across *data exposure* → *representation* → *generation*, covering pretraining and post-training data, tokenization and tokenizer provenance, model prediction cost, and generated language (see Figure 1). To our knowledge, this is the first systematic and rigorous pipeline-wide experimental study of *structural bias* across the LLM development pipeline.

This pipeline begins with data. Modern LLMs are trained on massive internet-derived corpora (Dodge et al., 2021), and English accounts for roughly 50–60% of global web content (Petrosyan, 2025). Its share is even higher in several documented training mixtures: approximately 92.65% for GPT-3,[1] 89.7% for Llama 2 (Touvron et al., 2023), and nearly 90% for Claude 2 (Anthropic, 2023). Yet the English-language Internet is itself uneven. The global reach of U.S. media, technology companies, commerce, and digital platforms has made AmE especially prominent online (Gonçalves et al., 2018). Because web-scale pretraining draws from this ecosystem, inequalities in which English forms are produced and circulated online can become inequalities in model training. These imbalances may extend beyond linguistic form, as dominant sources also encode cultural conventions, institutional practices, and perspectives that receive greater visibility in training data (Bender et al., 2021).

We examine a wider infrastructure question: how globally uneven data ecosystems and historically shaped linguistic standards become reflected in foundation-model infrastructure. To measure this asymmetry, we use British English as a controlled reference, denoted $\mathrm{Ref}_{\mathrm{BrE}}$. AmE and BrE provide well-documented contrasts across orthography, vocabulary, grammar, writing conventions, collocations, and multi-word usage.[2] The comparison is also globally meaningful. Through histories of British colonial expansion, BrE conventions became institutionally embedded in education, government, journalism, and law across many English-using contexts, including India, Nigeria, Australia, and New Zealand, among many others (Crystal, 2003; Tikly, 2016). Later U.S. economic, cultural, technological, and digital influence expanded the international reach of AmE (Calabrese et al., 2015; Gonçalves et al., 2018). These matched regional forms therefore provide a measurable lens for evaluating whether AmE receives a consistent advantage across the LLM pipeline.

Methodologically, we ground the analysis in $1,813$ matched AmE–$\mathrm{Ref}_{\mathrm{BrE}}$ variants and introduce DIALIGN, a dynamic, training-free method for estimating regional alignment from lexical, grammatical, structural, stylistic, and multi-word evidence (§2.1). We evaluate three stages: ① **Data exposure**, across six pretraining corpora and 21 public post-training datasets (§3); ② **Representation**, through

[1] OpenAI GPT-3 Dataset Language Statistics (GitHub, accessed August 06, 2026)

[2] Representative distinctions between AmE and BrE are shown in Table 6, and more in Tables 18 and 19.

tokenization, tokenizer provenance, semantic-equivalence controls, and model prediction cost (§4); and ③ **Generation**, under neutral English and explicit British-English prompting (§5). The evidence follows the same direction across all three stages. AmE is favored in all six audited pretraining corpora and all 21 post-training datasets; its forms are generally tokenized more compactly; matched $\mathrm{Ref}_{\mathrm{BrE}}$ counterfactuals receive higher per-token prediction cost across all ten evaluated checkpoints under the strict equal-token-count control; and AmE remains the dominant generation default under neutral English prompting. $\mathrm{Ref}_{\mathrm{BrE}}$ prompting shifts this preference, but does not consistently eliminate the AmE default. The downstream preference is therefore accompanied by upstream asymmetries in data exposure and representation, revealing a pipeline-wide pattern. The framework also identifies where component-level interventions can be directed, motivating practical recommendations for corpus construction and filtering, tokenizer design and provenance auditing, post-training data selection, and model evaluation (§F). We organize the investigation around three core research questions.

Research Questions

RQ1: Data Exposure. *Where does an American English advantage appear in the data used to train and align LLMs?* We audit pretraining and post-training data to measure AmE representation (§3).

RQ2: Representation. *Does the AmE data advantage persist in tokenization and prediction cost?* We examine tokenization, provenance, semantic-equivalence controls, and paired model likelihood (§4).

RQ3: Generation. *When LLMs generate English, does American English emerge as the default?* We compare neutral English and British-English prompting across different datasets and registers (§5).

2 Methodology: Resources and Alignment

We introduce DiAlign, a dynamic, training-free method for estimating regional alignment from contrastive distributional evidence. The method compares a study variety against a reference distribution; here, American English (AmE) is the study variety and British English the controlled reference, Ref(BrE). Rather than scoring documents by topic or style, it decomposes text into overlapping local spans and aggregates lexical, grammatical, structural, stylistic, and multi-word evidence. The procedure generalizes to other variety pairs with comparable reference distributions (Appendix B.6).

Regional Variant Corpus. We construct a curated corpus of 1,813 matched AmE–Ref(BrE) variants for consistent analysis across the pipeline. The pairs were manually compiled from authentic linguistic references and web-based lexicons (Table 17 of Appendix) and cover orthographic and lexical differences. Construction, filtering, and categorization details are provided in Appendix A.

2.1 DiAlign: Regional Alignment Score

Given a tokenized input $x = (t_1, \ldots, t_N)$ of length N, DiAlign returns $(P_{\mathrm{AmE}}, P_{\mathrm{Ref}_{\mathrm{BrE}}})$, representing relative alignment with the study and reference varieties. The procedure has four stages.

n-gram Extraction. We extract all contiguous n-grams from x for $2 \le n \le 5$:

$$\mathcal{G}(x) = \bigcup_{n=2}^{5} \{g = (t_i, \ldots, t_{i+n-1}) \mid 1 \le i \le N - n + 1\},$$

$$|\mathcal{G}(x)| = \sum_{n=2}^{5} (N - n + 1) = 4N - 10, \qquad N \ge 5.$$

Candidate spans are then filtered, providing dense local evidence even for short inputs. To reduce topical and non-diagnostic signals, we discard spans $g \in \mathcal{G}(x)$ containing named entities or consisting entirely of stopwords. Additional filtering details are provided in Appendix B.1.

Frequency Lookup. For each retained span g, we query separate American and British English corpora in Google Books Ngram[3] and obtain normalized average yearly frequencies $f_{\mathrm{AmE}}(g)$ and $f_{\mathrm{Ref}_{\mathrm{BrE}}}(g)$ over $[y_{\min}, y_{\max}]$. Thus, the same local construction is compared across two variety-specific frequency distributions under a common scale, rather than scored by raw frequency in a single corpus. Spans with zero frequency in either reference are discarded.

[3] https://books.google.com/ngrams/.

Signed Divergence. For each retained span, we compute the signed log-frequency ratio:

$$\mathrm{LR}(g) = \log_2\left(\frac{f_{\mathrm{AmE}}(g)}{f_{\mathrm{Ref_{BrE}}}(g)}\right).$$

Positive values favor AmE, negative values favor Ref(BrE), and $\mathrm{LR}(g) = 0$ indicates no directional signal. We down-weight weakly differentiated spans using $\delta(g)$ and give additional weight to spans containing an entry from the curated variant lexicon $\mathcal{D}$ using $w(g)$:

$$\delta(g) = \frac{|f_{\mathrm{AmE}}(g) - f_{\mathrm{Ref_{BrE}}}(g)|}{f_{\mathrm{AmE}}(g) + f_{\mathrm{Ref_{BrE}}}(g)} \in [0,1), \qquad w(g) = \begin{cases} \delta(g)\beta, & g \cap \mathcal{D} \neq \emptyset, \\ \delta(g), & \text{otherwise.} \end{cases}$$

Here, $\beta > 1$ increases the contribution of explicitly variety-diagnostic evidence.

Aggregation and Normalization. We aggregate positive and negative evidence separately:

$$S_{\mathrm{AmE}} = \sum_{\substack{g \in \mathcal{G}(x) \\ \mathrm{LR}(g) > 0}} \mathrm{LR}(g)\, w(g), \qquad S_{\mathrm{Ref_{BrE}}} = \sum_{\substack{g \in \mathcal{G}(x) \\ \mathrm{LR}(g) < 0}} |\mathrm{LR}(g)|\, w(g).$$

We then normalize:

$$P_{\mathrm{AmE}} = \frac{S_{\mathrm{AmE}}}{S_{\mathrm{AmE}} + S_{\mathrm{Ref_{BrE}}}}, \qquad P_{\mathrm{Ref_{BrE}}} = \frac{S_{\mathrm{Ref_{BrE}}}}{S_{\mathrm{AmE}} + S_{\mathrm{Ref_{BrE}}}}.$$

The predicted alignment is

$$\hat{y}(x) = \arg \max_{d \in \{\mathrm{AmE}, \mathrm{Ref_{BrE}}\}} P_d.$$

Meta-evaluation. We evaluate DIALIGN on a balanced benchmark of 1,500 news passages: 750 American English texts from HuffPost U.S. News and 750 British English texts from BBC England, using Google Books reference frequencies from 1950–2022. DIALIGN achieves **93.18% accuracy**, 90.67% precision, 96.25% recall, and 93.38 F1 (Figure 2). Correct predictions concentrate at larger confidence margins, while most errors occur near the decision boundary. Restricting the reference window to 2000–2022 yields similar performance, with **92.91% accuracy** and 93.12 F1, indicating temporal robustness. Benchmark construction, ablations, confidence analysis, and full temporal-robustness results are reported in Appendix B.3.

Implementation details are provided in Appendix B.1, with a span-level walkthrough in Appendix B.2. Benchmark construction, full validation results, ablations, analyses, and temporal robustness are reported in Appendix B.3. Generalization beyond the AmE–Ref(BrE) setting is discussed in Appendix B.6.

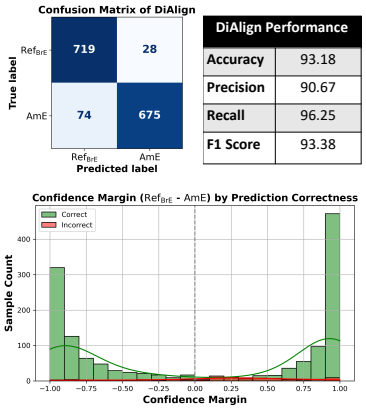


Figure 2: Meta-evaluation of DIALIGN on a balanced AmE–Ref(BrE) benchmark, showing the confusion matrix, classification performance, and confidence-margin distributions.

3 RQ1: DATA EXPOSURE ACROSS PRETRAINING AND POST-TRAINING

Our goal is to determine whether American English is systematically overrepresented in the data used to train and align LLMs. We examine two stages of data exposure: **(1)** six open-access pretraining corpora and **(2)** 21 public post-training datasets. Using the matched lexical variants from Section 2, we measure orthographic and vocabulary differences from corpus frequencies. For longer pretraining text, we also apply DIALIGN (§2.1) to capture broader grammatical, structural, stylistic, and multi-word variation. This audit establishes whether an AmE advantage is already visible in the data.

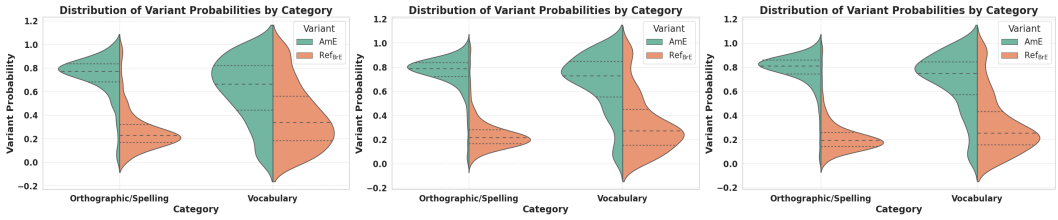
3.1 PRETRAINING DATA AUDIT

Methodology. For each lexical pair, we extract corpus frequencies f_{AmE} and $f_{\mathrm{Ref_{BrE}}}$ and compute

$$P_{\mathrm{AmE}} = \frac{f_{\mathrm{AmE}}}{f_{\mathrm{AmE}} + f_{\mathrm{Ref_{BrE}}}}, \qquad P_{\mathrm{Ref_{BrE}}} = \frac{f_{\mathrm{Ref_{BrE}}}}{f_{\mathrm{AmE}} + f_{\mathrm{Ref_{BrE}}}}. \tag{1}$$

Table 1: Variant usage across six pretraining corpora. Under the **Lexical Variants** analysis, we report corpus-level distributions for orthographic (spelling-based) and vocabulary-based AmE–Ref(BrE) differences. We also report **DIALIGN** scores, which capture broader regional alignment from grammatical, structural, collocational, and multi-word evidence; bracketed values denote mean alignment confidence. Datasets marked with * are evaluated on sampled subsets. RedPajama and Dolma contain mixed-domain content, including </> code, papers, forums, and social media. All probabilities are reported as percentages. Token statistics were computed using the LLaMA tokenizer (Grattafiori et al., 2024).

Data Source	Document Type	Documents (millions)	Tokens (billions)	Lexical Variants: Orthographic AmE	Lexical Variants: Orthographic Ref_{BrE}	Lexical Variants: Vocabulary AmE	Lexical Variants: Vocabulary Ref_{BrE}	DiAlign (broader variants): % of AmE	DiAlign (broader variants): % of Ref_{BrE}
Book Corpus (2015)	books	74	1.28	86.81	13.19	75.00	25.00	79.90 [0.79]	20.10 [0.76]
Wikipedia (2024)	encyclopedic	6.4	4.3	72.94	27.06	61.43	38.57	72.18 [0.80]	27.82 [0.78]
Common Crawl (C4) (2020)	web pages	365	156	75.12	24.88	67.00	33.00	73.67 [0.81]	26.33 [0.81]
Falcon RefinedWeb (2023)	web pages	968	600	77.34	22.66	68.35	31.65	74.04 [0.82]	25.96 [0.75]
RedPajama* (2024)	mixed	0.93	1.0	76.03	23.97	66.05	33.95	72.38 [0.85]	27.62 [0.75]
Dolma* (2024)	mixed	14.28	10	77.30	22.70	67.77	32.23	75.13 [0.81]	24.87 [0.77]



(a) Wikipedia (b) Common Crawl (C4) (c) Dolma

Figure 3: Distribution of AmE and Ref(BrE) variant probabilities across three pretraining corpora, shown separately for orthographic and vocabulary contrasts. Probabilities are computed from corpus frequencies over the matched variant pairs. Orthographic contrasts exhibit the strongest AmE skew, while vocabulary contrasts are more dispersed but lean toward AmE. Results for the remaining corpora are provided in Appendix Figure 7.

These pair-level probabilities measure the relative prevalence of matched forms within each corpus. We aggregate them by corpus and report orthographic and vocabulary variants separately for direct comparison. To test directional consistency, we apply the Wilcoxon signed-rank test to paired frequency differences $(f_{\text{AmE}} - f_{\text{Ref}_{\text{BrE}}})$, following standard practice for skewed corpus distributions (Dror et al., 2018). All six corpora differ significantly from parity ($p < 0.01$). Dataset and preprocessing details are provided in Appendices C.1 and C.2.

All six pretraining corpora favor AmE. Table 1 shows a consistent AmE skew across all audited corpora. The strongest asymmetry appears in orthographic variants, where AmE spellings often exceed 70% of observed pair frequencies. Vocabulary contrasts are more balanced but still lean toward AmE. DIALIGN shows the same direction for broader grammatical, structural, stylistic, and multi-word evidence, indicating that the pattern extends beyond isolated lexical choices.

Orthographic skew is stronger and more systematic than vocabulary skew. Figure 3 shows orthographic variants clustering more strongly toward AmE across the corpora, while vocabulary variants are more dispersed. The finer-grained breakdown in Appendix Figure 8 shows the same pattern across ten linguistic subcategories, including `-ize/-ise` and `-og/-ogue`. Most favor AmE; `-og/-ogue` is comparatively balanced, consistent with continued use of forms such as `catalogue` and `dialogue` in some American formal writing (Neumann, 2023).

3.2 POST-TRAINING DATA AUDIT

Post-training introduces another major source of linguistic exposure after pretraining through instruction tuning, chat demonstrations, human feedback, preference data, reward modeling, and safety alignment. We audit 21 widely used public datasets spanning these stages. Because many post-training samples are short, we use the explicit matched-variant probabilities as defined in Eq. 1.

All 21 post-training datasets favor AmE. Across `11,574,254` samples, we identify `28,860,119` explicit variant occurrences, of which **76.17%** are AmE. Every category in Table 2 favors AmE. Instruction-tuning and task-mixture data are 75.95% and 71.77% AmE, respectively, while chat, human-feedback, reward-model feedback, and safety-preference data reach 84.35%, 83.08%, 84.72%,

Table 3: Fertility scores for AmE and Ref(BrE) variants across the evaluated tokenizers, reported separately for orthographic and vocabulary contrasts. Lower fertility indicates more compact tokenization. Δ_o and Δ_v denote the relative fertility gap between matched AmE and Ref(BrE) forms. **Bold** and underlined values indicate the lowest and second-lowest fertility, respectively. All AmE–Ref(BrE) differences are statistically significant under the Wilcoxon signed-rank test ($p < 0.01$), except for Velvet-2B in the orthographic category, marked with †.

Tokenizers	Developer Country	Model Access	Vocab Size	Orthographic			Vocabulary		
				AmE (🇺🇸)	Ref_{BrE}(🇬🇧)	Δ_o	AmE (🇺🇸)	Ref_{BrE}(🇬🇧)	Δ_v
GPT-4	USA (🇺🇸)	🔒	100K	2.73	2.86	↑ 4.76 %	2.27	2.64	↑ 16.30 %
GPT-4o	USA (🇺🇸)	🔒	200K	2.65	2.77	↑ 4.53 %	2.21	2.57	↑ 16.29 %
Llama-3.3-70B	USA (🇺🇸)	🔓	128K	2.72	2.85	↑ 4.78 %	2.27	2.63	↑ 15.86 %
Gemma-3-27B	USA (🇺🇸)	🔓	262K	**2.40**	**2.53**	↑ 5.42 %	**2.02**	**2.35**	↑ 16.34 %
DeepSeek-V3	China (🇨🇳)	🔓	128K	2.71	2.80	↑ 3.32 %	2.37	2.67	↑ 12.66 %
Mistral-Small-24B	France (🇫🇷)	🔓	131K	2.81	2.89	↑ 2.85 %	2.45	2.79	↑ 13.88 %
StableLM-2-1.6B	UK (🇬🇧)	🔓	100K	2.73	2.86	↑ 4.76 %	2.27	2.64	↑ 16.30 %
Velvet-2B	Italy (🇮🇹)	🔓	127K	2.90	2.88	↓ 0.69 %†	2.40	2.72	↑ 13.33 %
Falcon3-7B	UAE (🇦🇪)	🔓	131K	2.44	2.56	↑ 4.92 %	2.03	2.41	↑ 18.72 %

and 86.86%. Preference, reward-modeling, and safety resources contain 4,666,867 variant occurrences and are **80.94%** AmE overall.

The AmE skew persists across datasets despite varying magnitude. All 21 datasets favor AmE, ranging from 65.29% in Databricks Dolly 15k to 89.72% in WizardLM Evol-Instruct v2 (Appendix 9). As in pretraining, orthographic contrasts show a stronger AmE preference than vocabulary contrasts. The same directional pattern across heterogeneous resources indicates that the skew is not confined to a single post-training dataset or data type.

AmE overrepresentation spans both data stages. AmE forms are favored in all six pretraining corpora and all 21 post-training datasets. Although the magnitude varies across resources, the direction remains stable, establishing data exposure as the first stage where the AmE advantage is systematically visible.

Table 2: **American English is consistently overrepresented across public post-training data.** Results aggregate 21 datasets spanning instruction tuning, human feedback, preference, reward modeling, and safety. **Variant hits** counts matches to either form in the curated AmE–Ref(BrE) pairs, and **AmE (%)** reports the weighted AmE share. Dataset-level results in Appendix Table 9.

Post-training stage	Hits	AmE (%)
Instruction tuning / SFT	16,348,769	75.95
Instruction tuning / chat	1,204,637	84.35
Instruction tuning / hum feedback	61,030	83.08
Instruction tuning / task mixture	6,578,816	71.77
RLHF / preference data	4,168,689	80.47
RLHF / reward-model feedback	77,332	84.72
RLHF / safety feedback	242,377	83.48
RLHF / safety preference	178,469	86.86
Preference / reward / safety	**4,666,867**	**80.94**
All post-training data	**28,860,119**	**76.17**

Key Takeaway: American English is clearly overrepresented across pretraining and post-training data, establishing an exposure asymmetry before downstream representation in RQ2 and generation in RQ3.

4 RQ2: Regional Representation and Prediction-Cost Asymmetry

Our aim is to determine whether the AmE advantage observed in data exposure is also visible in how LLMs *represent* and *predict* matched regional forms. We examine two complementary levels: whether tokenizers encode AmE variants more compactly than Ref_{BrE} counterparts, and whether language models assign lower prediction cost to AmE forms in otherwise equivalent contexts. This separates tokenizer-level representational asymmetry from preference expressed by the model itself.

4.1 Tokenization: Representation Efficiency

We examine how public tokenizers encode the matched AmE–Ref_{BrE} variants. Our primary measure is *fertility*, the average number of subword tokens required to encode a word (Rust et al., 2021). Lower fertility indicates more compact representation as averages can obscure fragmented forms, we also measure *granularity*: how often variants are represented by 1, 2, or 3+ subwords. Fertility captures average encoding cost, while granularity reveals its distribution across regional forms.

AmE forms are generally represented more compactly. Across the evaluated tokenizers, Table 3 and Figure 4 show that Ref_{BrE} variants require more subword tokens than matched AmE forms. The gap is largest for vocabulary contrasts, reaching $\Delta_v = 18.72\%$, while orthographic gaps

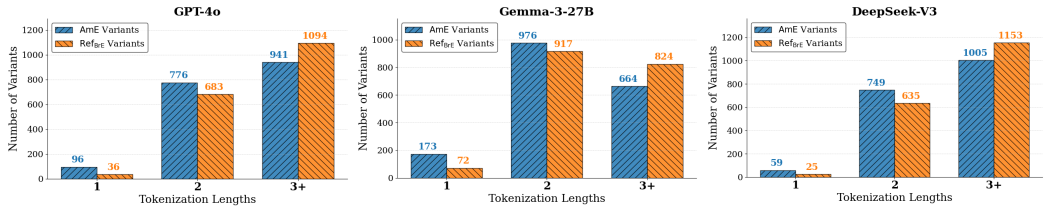


Figure 4: Tokenization granularity for AmE and $\mathrm{Ref}_{\mathrm{BrE}}$ variants across three tokenizers. Each panel shows the number of variants represented by 1, 2, or 3+ subwords. $\mathrm{Ref}_{\mathrm{BrE}}$ forms appear more frequently in the 3+ subword category, indicating greater long-tail fragmentation. Remaining tokenizers appear in Appendix Fig. 11.

among AmE-favoring tokenizers range from $\Delta_o = 2.85\%$ to 5.42%. Granularity shows the same pattern: $\mathrm{Ref}_{\mathrm{BrE}}$ forms occur more often in the 3+ token category, indicating greater long-tail fragmentation. The magnitude varies across tokenizers. Velvet slightly favors $\mathrm{Ref}_{\mathrm{BrE}}$ orthographic forms ($\Delta_o = -0.69\%$); DeepSeek has the smallest vocabulary gap ($\Delta_v = 12.66\%$); and Gemma has the lowest fertility for both forms, consistent with its larger 262K-token vocabulary. The pattern is therefore more compact representation of AmE, not uniform behavior across tokenizer families.

Tokenizer reuse can transmit representational behavior across model families. Tokenizer provenance provides context for these differences. Of the nine tokenizers evaluated, seven use independent or source vocabularies, while Llama-3.3 and StableLM-2 inherit or extend U.S.-origin tokenizer artifacts. StableLM-2 provides the clearest case: despite its association with a UK-developed model, its tokenizer shows **100% token-boundary identity** with GPT-4 on both the AmE and $\mathrm{Ref}_{\mathrm{BrE}}$ variant sets, with all `100,256` shared base tokens retaining identical rank/id assignments. We therefore separate tokenizer provenance from model developer location. Tokenizer reuse provides a concrete mechanism through which representational conventions can persist across distinct model families. Full vocabulary-overlap, rank-identity, and token-boundary analyses are reported in Appendix D.2.

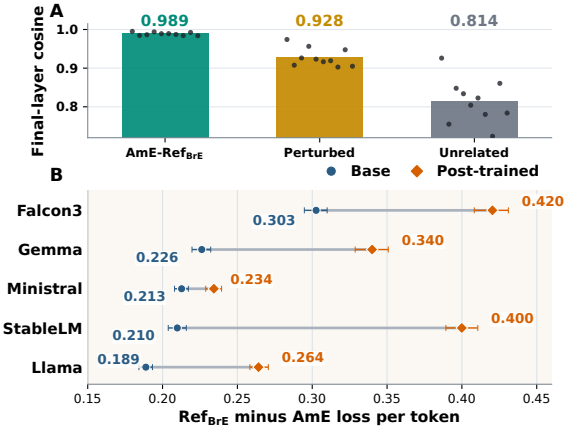


Figure 5: **Semantic equivalence, unequal prediction cost.** Top: matched AmE–$\mathrm{Ref}_{\mathrm{BrE}}$ counterfactuals are more similar than character-perturbed and unrelated controls. Bottom: $\mathrm{Ref}_{\mathrm{BrE}}$ counterfactuals receive higher per-token loss across all five base/post-trained model families under equal sentence token counts. Error bars show paired-bootstrap 95% CIs.

4.2 Prediction Cost: Model Preference Beyond Tokenization

Compact segmentation does not by itself show which regional form a model finds more probable. We therefore construct paired counterfactual sentences $(x_i^{\mathrm{AmE}}, x_i^{\mathrm{Ref}_{\mathrm{BrE}}})$ that differ only in the target regional form and compare prediction costs across five base/post-trained model pairs (10 checkpoints).

Methodology. For model m, we define per-token negative log-likelihood as

$$\ell_m(x) = -\frac{1}{|\tau_m(x)|} \sum_{t=1}^{|\tau_m(x)|} \log p_m(w_t \mid w_{<t}), \qquad \Delta_{m,i}^{\mathrm{loss}} = \ell_m(x_i^{\mathrm{Ref}_{\mathrm{BrE}}}) - \ell_m(x_i^{\mathrm{AmE}}).$$

Here, $\tau_m(x)$ is the model token sequence. Positive $\Delta_{m,i}^{\mathrm{loss}}$ indicates higher prediction cost for the $\mathrm{Ref}_{\mathrm{BrE}}$ counterfactual; for confirmatory analysis, we retain pairs with identical sentence token counts,

$$|\tau_m(x_i^{\mathrm{AmE}})| = |\tau_m(x_i^{\mathrm{Ref}_{\mathrm{BrE}}})|,$$

so the remaining loss difference cannot be attributed to unequal token counts.

We also verify semantic equivalence. Let $h_{m,\ell}(x)$ denote the mean-pooled contextual representation at layer ℓ. For each pair,

$$\mathrm{sim}_{m,\ell,i} = \cos\Big(h_{m,\ell}(x_i^{\mathrm{AmE}}), h_{m,\ell}(x_i^{\mathrm{Ref_{BrE}}})\Big),$$

which we compare against character-perturbed and unrelated controls. This serves as a validity check for the paired prediction-cost experiment.

Semantically equivalent forms receive systematically different prediction costs. Matched AmE–$\mathrm{Ref_{BrE}}$ counterfactuals are substantially closer in contextual representation space than character-perturbed or unrelated controls (Figure 5), supporting their semantic equivalence. Yet all ten evaluated checkpoints assign higher per-token loss to $\mathrm{Ref_{BrE}}$ on the strict equal-token-count slice, with all paired-bootstrap 95% confidence intervals excluding zero. Because sentence token count is held fixed, this prediction-cost gap is distinct from the fragmentation difference observed in tokenization. The AmE advantage therefore appears in both segmentation and model-assigned probability.

The prediction-cost gap is larger in every evaluated post-trained checkpoint. Across the five model families with matched public base and post-trained checkpoints, the $\mathrm{Ref_{BrE}}$ loss gap is larger after post-training in every family: Gemma, Ministral, Llama, StableLM, and Falcon3. This is a consistent checkpoint-level pattern; isolating a causal post-training contribution requires controlled intervention. Full results, bootstrap intervals, and controlled slices are provided in Appendix D.6; semantic-equivalence diagnostics in Appendix D.8; and model-selection details in Appendix D.7.

Key Takeaway: The AmE advantage appears at two representation levels: AmE forms are encoded more compactly, while semantically matched $\mathrm{Ref_{BrE}}$ forms incur higher prediction cost even at equal token counts. Tokenizer lineage shows how inherited vocabularies can carry representational behavior across model families.

5 RQ3: Regional Preference in Generation

Our goal is to determine whether the AmE advantage observed in data exposure (RQ1) and representation (RQ2) is reflected in generated language. We use DiAlign (§2.1) to measure regional alignment under neutral English prompting and explicit British-English prompting.

Experimental Setup. We evaluate open-domain question answering across two contrasting registers: Natural Questions (NQ) (Kwiatkowski et al., 2019) as a more formal, encyclopedic setting, and ELI5 (Fan et al., 2019) as a more informal, conversational setting. To avoid lexical priming, we remove questions containing variants from the AmE–Ref(BrE) lexicon. For each question, we generate responses under two conditions: `English` and `British English (en-GB)`, requesting 50 words to keep response length comparable across models and conditions for controlled comparison.

Table 4: Regional preferences in LLM generation on Natural Questions (*formal*) and ELI5 (*informal*). We report the percentage of outputs classified as AmE under default English and British English (en-GB) prompting, with mean AmE alignment scores in brackets. AmE is the dominant default, while British-English prompting generally shifts outputs toward Ref(BrE). **Bold** marks the lowest AmE percentage in each column; underlined marks the second-lowest.

LLMs	Developer Country	Natural Questions [formal]		ELI5 [informal]	
		Default English ($\mathrm{AmE^{Default}}$)	British English ($\mathrm{AmE^{BrE}}$)	Default English ($\mathrm{AmE^{Default}}$)	British English ($\mathrm{AmE^{BrE}}$)
GPT-4o	USA	79.00% [0.81]	45.33% [0.77]	77.00% [0.82]	34.67% [0.78]
Gemini-2.0-flash	USA	76.00% [0.83]	51.00% [0.79]	75.33% [0.82]	42.33% [0.80]
Claude-3.7-sonnet	USA	75.67% [0.85]	42.33% [0.80]	73.33% [0.86]	37.67% [0.78]
Llama-3.3-70B	USA	74.67% [0.82]	47.33% [0.78]	69.00% [0.79]	**30.00%** [0.76]
Gemma-3-27B	USA	**69.33%** [0.81]	45.67% [0.78]	68.33% [0.83]	38.00% [0.78]
DeepSeek-V3	China	74.67% [0.82]	**41.67%** [0.78]	73.33% [0.84]	40.47% [0.76]
Mistral-Small-24B	France	73.67% [0.81]	48.00% [0.75]	72.33% [0.82]	38.67% [0.76]
StableLM-2-1.6B	UK	74.00% [0.79]	69.67% [0.78]	73.33% [0.77]	67.00% [0.77]
Velvet-2B	Italy	72.91% [0.80]	69.33% [0.81]	71.00% [0.80]	67.67% [0.80]
Falcon3-7B	UAE	73.67% [0.80]	66.00% [0.79]	**66.00%** [0.81]	63.33% [0.77]

Even at this length, overlapping 2–5-grams provide substantial local evidence for DiAlign (§2.1). Alignment is also effectively uncorrelated with response length ($r = 0.017$ overall; Appendix E). For each response, DiAlign returns $(P_{\mathrm{AmE}}, P_{\mathrm{Ref_{BrE}}})$. Table 4 reports the percentage of generations classified as AmE and mean AmE alignment P_{AmE} in brackets, denoted $\mathrm{AmE^{Default}}$ and $\mathrm{AmE^{BrE}}$ for the two prompting conditions. Dataset filtering, prompts (Figure 13), model details, decoding parameters, and length-robustness analyses are provided in Appendix E.

AmE is the dominant generation default under neutral English prompting. Table 4 shows a consistent AmE preference when models are instructed to respond in English. Across most models

and both datasets, a majority of outputs are classified as AmE, often with strong alignment scores. The upstream asymmetries identified in RQ1 and RQ2 therefore also appear in realized behavior: without regional steering, generated responses align more strongly with AmE than with $\mathrm{Ref}_{\mathrm{BrE}}$.

British-English prompting shifts, but does not consistently eliminate, the AmE default. Requesting `British English (en-GB)` shifts outputs toward $\mathrm{Ref}_{\mathrm{BrE}}$ across models, but substantial AmE alignment often remains, particularly in the more formal NQ setting. The magnitude of this shift varies across conditions, models, and registers, showing that downstream prompting can alter realized language without consistently removing the AmE default. Such output-level steering does not alter the upstream asymmetries already empirically established in RQ1 and RQ2.

The strength of the AmE default varies across models and registers. The AmE preference is generally stronger in NQ than ELI5, while British-English prompting produces larger shifts for several models in the more conversational ELI5 setting. For example, Llama-3.3 falls to about 30% AmE-classified outputs on ELI5 under British-English prompting. The default is therefore not uniform: its magnitude depends on both the specific model and local linguistic context.

Key Takeaway: American English is the dominant generation default under neutral English prompting. British-English prompting shifts outputs toward $\mathrm{Ref}_{\mathrm{BrE}}$ but does not consistently eliminate the AmE default, with the magnitude of the shift varying across models and registers. RQ3 provides clear downstream behavioral evidence for the AmE advantage already established in data exposure (RQ1) and representation (RQ2).

6 Related Work

Training data audits and curation. Prior work examines how dataset composition, quality, provenance, and curation shape model behavior. Studies show that data age, coverage, and quality affect generalization (Longpre et al., 2024), while audits identify duplication, contamination, and provenance gaps in widely used corpora (Elazar et al., 2024; Longpre et al., 2025). Other work studies corpus construction (Parmar et al., 2024), register- and domain-aware sampling (Myntti et al., 2025), and reuse of filtered web data (Nguyen et al., 2025). We add within-English variation to this auditing agenda by measuring AmE representation across pretraining and post-training data.

Tokenization and representation. Tokenization can encode systematic differences before generation begins. Prior work documents substantial variation in tokenization efficiency across languages, with consequences for sequence length, context capacity, and inference cost (Petrov et al., 2023; Rust et al., 2021; Ahia et al., 2023; Ali et al., 2024). Related studies connect tokenization to demographic bias (An & Rudinger, 2023), lexical fragility (Gallifant et al., 2024), and gender bias in machine translation (Iluz et al., 2023). We study this problem within English by combining tokenization efficiency, tokenizer provenance, and paired model likelihood for matched AmE and Ref(BrE) forms.

English variation in NLP and LLMs. Research on African American English (AAE), South Asian Englishes (SAsE), and other English varieties has shown systematic disparities in tagging, classification, sentiment analysis, robustness, and model behavior (Jørgensen et al., 2016; Blodgett et al., 2016; Kiritchenko & Mohammad, 2018; Ziems et al., 2023; Hofmann et al., 2024; Fleisig et al., 2024; Holt et al., 2024). We focus on a different question: how an American-English default appears across the LLM pipeline. By linking training data, tokenization and prediction cost, and generation, our analysis treats regional preference as a structural property that can be observed across multiple stages of model development. An extended discussion is provided in Appendix G.

7 Conclusion

This paper presents the first rigorous pipeline-wide study of *structural bias* across phases of LLM development. By triangulating pretraining and post-training data, tokenizer behavior and provenance, model prediction cost, and generated language, we show that American English receives an advantage across data exposure, representation, and generation, emerging as the *de facto* default. These findings show how linguistic asymmetries persist across the LLM pipeline, raising concerns about linguistic homogenization, epistemic injustice, and inequity in global AI deployment, while providing a basis for targeted, component-level improvements toward more inclusive language technologies.

AI USE STATEMENT

We used generative AI tools in limited assistive roles. Specifically, they were used for **(1) text editing**, including improvements to grammar, spelling, phrasing, and word choice; **(2) code assistance**, including code fragments for data preprocessing and filtering, routine implementation tasks, and debugging; and **(3) visual assistance**, limited to generating decorative icons used in Figure 1.

All AI-assisted text, code, and visual materials were reviewed, adapted, and approved by the authors, and any generated or modified code was manually verified and tested before use. The authors developed the research questions and methodology, designed the experiments, conducted the analyses, interpreted the results, and determined the final presentation and conclusions. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies structural linguistic bias across the LLM development pipeline. It involves no human subjects, private user data, or sensitive personal attributes. Our analyses use publicly available pretraining and post-training datasets for research purposes (Appendix C.1). The curated AmE–$\mathrm{Ref}_{\mathrm{BrE}}$ resource was compiled from publicly accessible linguistic and web-based sources (Table 17) and contains no personal or proprietary material. We plan to release it for non-commercial research under a CC BY-NC-SA 4.0 license.[4]

The ethical motivation is to understand where linguistic advantage becomes visible, persists, and manifests across data exposure, representation, and generation. Such defaults can shape which regional forms receive greater visibility as LLMs are deployed across diverse national and institutional contexts, with implications for linguistic homogenization and unequal representation. British English is used solely as a controlled reference for measuring the American English preference; neither variety is treated as intrinsically superior, more correct, or more appropriate. Developer country and tokenizer provenance are likewise analyzed descriptively and are not interpreted as evidence of intent, nationality, or regional causation.

The pipeline-wide audit and DIALIGN are intended for responsible regional-language auditing and for identifying where targeted improvements may be most effective. DIALIGN excludes spans containing named entities such as people, organizations, and locations, reducing privacy and reputational risks from such signals. The framework can be extended beyond AmE–Ref(BrE) when suitable regional reference distributions and linguistic resources are available. Scope, limitations, and risks of binary regional comparison are discussed in Appendix I.

REPRODUCIBILITY STATEMENT

The paper and Appendix document the regional-variant resource, corpus construction and preprocessing, DIALIGN implementation and meta-evaluation, tokenizer provenance analysis, prediction-cost experiments, generation settings, prompts, evaluation metrics, and robustness analyses. All code, preprocessing scripts, resources, and experimental pipelines are available at **anonymous.4open.science/r/LLM-structural-bias**, including code for corpus auditing, DIALIGN, tokenization and provenance analysis, prediction-cost evaluation, generation scoring, and result analysis. The current release is prepared for anonymous review; a polished and fully documented version will be released upon the paper's acceptance.

REFERENCES

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. Do all languages cost the same? tokenization in the era of commercial language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9904–9923, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.614. URL https://aclanthology.org/2023.emnlp-main.614/.

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

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Buschhoff, Charvi Jain, Alexander Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, Stefan Kesselheim, and Nicolas Flores-Herr. Tokenizer choice for LLM training: Negligible or crucial? In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3907–3924, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.247. URL https://aclanthology.org/2024.findings-naacl.247/.

Sawsan Alqahtani, Mir Tafseer Nayeem, Md Tahmid Rahman Laskar, Tasnim Mohiuddin, and M Saiful Bari. Stop taking tokenizers for granted: They are core design decisions in large language models. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8410–8432, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.394. URL https://aclanthology.org/2026.eacl-long.394/.

Haozhe An and Rachel Rudinger. Nichelle and nancy: The influence of demographic attributes and tokenization length on first name biases. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 388–401, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.34. URL https://aclanthology.org/2023.acl-short.34/.

Anthropic. Model card and evaluations for claude models. https://www-cdn.anthropic.com/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226/Model-Card-Claude-2.pdf, 2023. Accessed September 27, 2025.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Paul Baker. *American and British English: Divided by a Common Language?* Cambridge University Press, Cambridge, 2017. ISBN 9781107088863. URL https://www.cambridge.org/9781107088863. Corpus-based analysis of historical and contemporary variation between British and American English.

Shabnam Behzad, Amir Zeldes, and Nathan Schneider. To ask LLMs about English grammaticality, prompt them in a different language. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15622–15634, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.916. URL https://aclanthology.org/2024.findings-emnlp.916/.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic dialectal variation in social media: A case study of African-American English. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1119–1130, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1120. URL https://aclanthology.org/D16-1120/.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. Twitter Universal Dependency parsing for African-American and mainstream American English. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), pp. 1415–1425, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1131. URL https://aclanthology.org/P18-1131/.

Rita Calabrese, J. K. Chambers, and Gerhard Leitner (eds.). *Variation and Change in Postcolonial Contexts*. Cambridge Scholars Publishing, Newcastle upon Tyne, UK, 1 edition, 2015. ISBN 978-1-4438-7611-7. URL https://www.cambridgescholars.com/product/978-1-4438-7611-7.

Andreas Chari, Sean MacAvaney, and Iadh Ounis. On the effects of regional spelling conventions in retrieval models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pp. 2220–2224, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394086. doi: 10.1145/3539618.3592030. URL https://doi.org/10.1145/3539618.3592030.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://lmsys.org/blog/2023-03-30-vicuna/, March 2023. Dataset used from https://huggingface.co/datasets/Aeala/ShareGPT_Vicuna_unfiltered. Accessed: September 11, 2026.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm. Accessed: September 11, 2026.

David Crystal. *English as a Global Language*. Cambridge University Press, Cambridge, UK, 2 edition, 2003. ISBN 9780521530323. doi: 10.1017/CBO9780511486999. URL https://www.cambridge.org/core/books/english-as-a-global-language/690D300092A3B5E4733677FCC9A42E98. Second edition with updated discussions on the global spread of English and its sociolinguistic implications.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. ULTRAFEEDBACK: Boosting language models with scaled AI feedback. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9722–9744. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/cui24f.html.

Dorottya Demszky, Devyani Sharma, Jonathan Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. Learning to recognize dialect features. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2315–2338, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.184. URL https://aclanthology.org/2021.naacl-main.184/.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3029–3051, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.183. URL https://aclanthology.org/2023.emnlp-main.183/.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1286–1305, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.98. URL https://aclanthology.org/2021.emnlp-main.98/.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1383–1392, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1128. URL https://aclanthology.org/P18-1128/.

Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A. Smith, and Jesse Dodge. What's in my big data? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RvfPnOkPV4.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5988–6008. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ethayarajh22a.html.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: Long form question answering. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3558–3567, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1346. URL https://aclanthology.org/P19-1346/.

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13541–13564, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.750. URL https://aclanthology.org/2024.emnlp-main.750/.

Negar Foroutan, Clara Meister, Debjit Paul, Joel Niklaus, Sina Ahmadi, Antoine Bosselut, and Rico Sennrich. Parity-aware byte-pair encoding: Improving cross-lingual fairness in tokenization, 2025. URL https://arxiv.org/abs/2508.04796.

Wikimedia Foundation. Wikimedia downloads – wikipedia dump archive, 2024. URL https://dumps.wikimedia.org. Accessed: February 11, 2026.

Jack Gallifant, Shan Chen, Pedro José Ferreira Moreira, Nikolaj Munch, Mingye Gao, Jackson Pond, Leo Anthony Celi, Hugo Aerts, Thomas Hartvigsen, and Danielle Bitterman. Language models are surprisingly fragile to drug names in biomedical benchmarks. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 12448–12465, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.726. URL https://aclanthology.org/2024.findings-emnlp.726/.

Bruno Gonçalves, Lucía Loureiro-Porto, José J. Ramasco, and David Sánchez. Mapping the americanization of english in space and time. *PLOS ONE*, 13(5):1–15, 05 2018. doi: 10.1371/journal.pone.0197741. URL https://doi.org/10.1371/journal.pone.0197741.

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 542 others. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4693–4703, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.413. URL https://aclanthology.org/2021.findings-acl.413/.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. Ai generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028):147–154, September 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07856-5. URL https://doi.org/10.1038/s41586-024-07856-5.

Faye Holt, William Held, and Diyi Yang. Perceptions of language technology failures from South Asian English speakers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4067–4081, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.241. URL https://aclanthology.org/2024.findings-acl.241/.

Bar Iluz, Tomasz Limisiewicz, Gabriel Stanovsky, and David Mareček. Exploring the impact of training data distribution and subword tokenization on gender bias in machine translation. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi (eds.), *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 885–896, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.57. URL https://aclanthology.org/2023.ijcnlp-main.57/.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of LLM via a human-preference dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=g0QovXbFw3.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Alex Qiu, Jiayi Zhou, Kaile Wang, Boxun Li, Sirui Han, Yike Guo, and Yaodong Yang. PKU-SafeRLHF: Towards multi-level safety alignment for LLMs with human preference. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31983–32016, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1544. URL https://aclanthology.org/2025.acl-long.1544/.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. Learning a POS tagger for AAVE-like language. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1115–1120, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1130. URL https://aclanthology.org/N16-1130/.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. Natural language processing for dialects of a language: A survey. *ACM Comput. Surv.*, 57(6), February 2025. ISSN 0360-0300. doi: 10.1145/3712060. URL https://doi.org/10.1145/3712060.

Svetlana Kiritchenko and Saif Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. In Malvina Nissim, Jonathan Berant, and Alessandro Lenci (eds.), *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 43–53, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2005. URL https://aclanthology.org/S18-2005/.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Minh Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Alexandrovich Glushkov, Arnav Varma Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Julian Mattick. Openassistant conversations - democratizing large language model alignment. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=VSJotgbPHF.

Abhinandan Kulal, Habeeb Ur Rahiman, Harinakshi Suvarna, N. Abhishek, and Sahana Dinesh. Enhancing public service delivery efficiency: Exploring the impact of ai. *Journal of Open Innovation: Technology, Market, and Complexity*, 10(3):100329, 2024. ISSN 2199-8531.

doi: https://doi.org/10.1016/j.joitmc.2024.100329. URL https://www.sciencedirect.com/science/article/pii/S2199853124001239.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

William Labov. *Sociolinguistic Patterns*. Conduct & communication. University of Pennsylvania Press, Philadelphia, 1972. ISBN 9780812286571. URL https://books.google.ca/books?id=7_gByAEACAAJ.

William Labov. *The Social Stratification of English in New York City*. Cambridge University Press, Cambridge, 2 edition, 2006. ISBN 9780521528054. URL https://www.cambridge.org/9780521528054. Second edition with updated findings on sociolinguistic variation and methodology.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S. Yu. Large language models in law: A survey. *AI Open*, 5:181–196, 2024. ISSN 2666-6510. doi: https://doi.org/10.1016/j.aiopen.2024.09.002. URL https://www.sciencedirect.com/science/article/pii/S2666651024000172.

Pietro Lesci, Clara Meister, Thomas Hofmann, Andreas Vlachos, and Tiago Pimentel. Causal estimation of tokenisation bias. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28325–28340, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1374. URL https://aclanthology.org/2025.acl-long.1374/.

Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. Syntactic annotations for the Google Books NGram corpus. In Min Zhang (ed.), *Proceedings of the ACL 2012 System Demonstrations*, pp. 169–174, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/P12-3029/.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22631–22648. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/longpre23a.html.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3245–3276, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.179. URL https://aclanthology.org/2024.naacl-long.179/.

Shayne Longpre, Nikhil Singh, Manuel Cherep, Kushagra Tiwary, Joanna Materzynska, William Brannon, Robert Mahari, Naana Obeng-Marnu, Manan Dey, Mohammed Hamdy, Nayan Saxena, Ahmad Mustafa Anis, Emad A. Alghamdi, Vu Minh Chien, Da Yin, Kun Qian, Yizhi LI, Minnie Liang, An Dinh, Shrestha Mohanty, Deividas Mataciunas, Tobin South, Jianguo Zhang, Ariel N. Lee, Campbell S. Lund, Christopher Klamm, Damien Sileo, Diganta Misra, Enrico Shippole, Kevin Klyman, Lester James Validad Miranda, Niklas Muennighoff, Seonghyeon Ye, Seungone Kim, Vipul Gupta, Vivek Sharma, Xuhui Zhou, Caiming Xiong, Luis Villa, Stella Biderman, Alex Pentland, Sara Hooker, and Jad Kabbara. Bridging the data provenance gap across text, speech, and video. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=G5DziesYxL.

Marc Marone and Benjamin Van Durme. Data portraits: Recording foundation model training data. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=ZrNRBmOzwE.

Rishabh Misra. News category dataset, 2022. URL https://arxiv.org/abs/2209.11429.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4, 2023. URL https://arxiv.org/abs/2306.02707.

Amanda Myntti, Erik Henriksson, Veronika Laippala, and Sampo Pyysalo. Register always matters: Analysis of LLM pretraining data through the lens of language variation. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=FqXXtSZWEZ.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback, 2022. URL https://arxiv.org/abs/2112.09332.

Mir Tafseer Nayeem, Sawsan Alqahtani, Md Tahmid Rahman Laskar, Tasnim Mohiuddin, and M Saiful Bari. Beyond fertility: Analyzing strr as a metric for multilingual tokenization evaluation, 2025. URL https://arxiv.org/abs/2510.09947.

Leo Neumann. Catalogue or catalog – british vs. american english. https://www.bachelorprint.com/au/british-english-vs-american-english/catalogue-or-catalog/, September 2023. Accessed February 05, 2026.

Thao Nguyen, Yang Li, Olga Golovneva, Luke Zettlemoyer, Sewoong Oh, Ludwig Schmidt, and Xian Li. Recycling the web: A method to enhance pre-training data quality and quantity for language models. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=lkjhBdz3rn.

Elizabeth Nielsen, Christo Kirov, and Brian Roark. Spelling convention sensitivity in neural language models. In Andreas Vlachos and Isabelle Augenstein (eds.), *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 1334–1346, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.98. URL https://aclanthology.org/2023.findings-eacl.98/.

Richard Nordquist. English language: History, definition, and examples. https://www.thoughtco.com/what-is-the-english-language-1690652, July 2024. Accessed October 25, 2025.

OECD. Digital government outlook 2026: From foundations to transformational impact. Technical report, OECD Publishing, Paris, 2026. URL https://doi.org/10.1787/0496b2bc-en. Accessed: 2026-09-15.

Jupinder Parmar, Shrimai Prabhumoye, Joseph Jennings, Bo Liu, Aastha Jhunjhunwala, Zhilin Wang, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Data, data everywhere: A guide for pretraining dataset construction. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10671–10695, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.596. URL https://aclanthology.org/2024.emnlp-main.596/.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon LLM: Outperforming curated corpora with web data only. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=kM5eGcdCzq.

Ani Petrosyan. Languages most frequently used for web content as of february 2025, by share of websites. https://www.statista.com/statistics/262946/most-common-languages-on-the-internet/, February 2025. Accessed February 26, 2026.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=78yDLKi95p.

Robert Phillipson. *Linguistic Imperialism*, pp. 1–7. John Wiley & Sons, Ltd, 2018. ISBN 9781405198431. doi: https://doi.org/10.1002/9781405198431.wbeal0718.pub2. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781405198431.wbeal0718.pub2.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), January 2020. ISSN 1532-4435. URL https://dl.acm.org/doi/abs/10.5555/3455716.3455856.

Nazneen Rajani, Lewis Tunstall, Edward Beeching, Nathan Lambert, Alexander M. Rush, and Thomas Wolf. No robots. https://huggingface.co/datasets/HuggingFaceH4/no_robots, 2023. Accessed: September 11, 2026.

Anka Reuel, Benjamin Bucknall, Stephen Casper, Timothy Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, Markus Anderljung, Ben Garfinkel, Lennart Heim, Andrew Trask, Gabriel Mukobi, Rylan Schaeffer, Mauricio Baker, Sara Hooker, Irene Solaiman, Sasha Luccioni, Nitarshan Rajkumar, Nicolas Moës, Jeffrey Ladish, David Bau, Paul Bricman, Neel Guha, Jessica Newman, Yoshua Bengio, Tobin South, Alex Pentland, Sanmi Koyejo, Mykel Kochenderfer, and Robert Trager. Open problems in technical AI governance. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=1nO4qFMiS0. Survey Certification.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3118–3135, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.243. URL https://aclanthology.org/2021.acl-long.243/.

Sha Sajadieh, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Lapo Santarlasci, Juan Pava, Nestor Maslej, Russ Altman, Erik Brynjolfsson, Carla Brodley, Jack Clark, Virginia Dignum, Vipin Kumar, James Landay, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Elham Tabassi, Russell Wald, Toby Walsh, and Dan Weld. Artificial intelligence index report 2026, 2026. URL https://arxiv.org/abs/2606.15708.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1163. URL https://aclanthology.org/P19-1163/.

Tariq Shahzad, Tehseen Mazhar, Muhammad Usman Tariq, Wasim Ahmad, Khmaies Ouahada, and Habib Hamam. A comprehensive review of large language models: Issues and solutions in learning environments. *Discover Sustainability*, 6(1):27, 2025. ISSN 2662-9984. doi: 10.1007/s43621-025-00815-8. URL https://doi.org/10.1007/s43621-025-00815-8.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.840. URL https://aclanthology.org/2024.acl-long.840/.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023. Accessed: September 11, 2026.

Atula Tejaswi, Nilesh Gupta, and Eunsol Choi. Exploring design choices for building language-specific LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10485–10500, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.614. URL https://aclanthology.org/2024.findings-emnlp.614/.

Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants, 2023. URL https://huggingface.co/datasets/teknium/OpenHermes-2.5. Accessed: September 11, 2026.

Leon Tikly. Language-in-education policy in low-income, postcolonial contexts: Towards a social justice approach. *Comparative Education*, 52(3):408–425, 2016. ISSN 0305-0068. doi: 10.1080/03050068.2016.1185272. URL https://doi.org/10.1080/03050068.2016.1185272.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5085–5109, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.340. URL https://aclanthology.org/2022.emnlp-main.340/.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer 2: Open-source dataset for training top-performing reward models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=PvVKUFhaNy.

Maurice Weber, Daniel Y Fu, Quentin Gregory Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Re, Irina Rish,

and Ce Zhang. Redpajama: an open dataset for training large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=lnuXaRpwvw.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. Qurating: selecting high-quality data for training language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. Organize the web: Constructing domains enhances pre-training data curation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=boSqwdvJVC.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=CfXh93NDgH.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. Lima: Less is more for alignment. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 55006–55021. Curran Associates, Inc., 2023. doi: 10.52202/075280-2400. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ac662d74829e4407ce1d126477f4a03a-Paper-Conference.pdf.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, Karthik Ganesan, Wei-Lin Chiang, Jian Zhang, and Jiantao Jiao. Starling-7b: Improving helpfulness and harmlessness with RLAIF. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=GqDntYTTbk.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pp. 19–27, USA, 2015. IEEE Computer Society. ISBN 9781467383912. doi: 10.1109/ICCV.2015.11. URL https://doi.org/10.1109/ICCV.2015.11.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. VALUE: Understanding dialect disparity in NLU. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3701–3720, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.258. URL https://aclanthology.org/2022.acl-long.258/.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. Multi-VALUE: A framework for cross-dialectal English NLP. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 744–768, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.44. URL https://aclanthology.org/2023.acl-long.44/.

Supplementary Material: Appendices

Appendix Contents

A Regional Variant Resource
- A.1 Construction and Filtering
- A.2 Variant Grouping and Typology
- A.3 Sources and Extended Examples

B DIALIGN: Implementation, Validation, and Generalization
- B.1 Implementation Details
- B.2 Illustrative Walkthrough
- B.3 Meta-evaluation Benchmark
- B.4 Validation and Ablation
- B.5 Temporal Robustness
- B.6 Generalizing DIALIGN

C Additional Details for RQ1: Data Exposure
- C.1 Pretraining Datasets
- C.2 Pretraining Preprocessing
- C.3 Corpus Processing and Sampling
- C.4 Additional Pretraining Results
- C.5 Post-Training Dataset Audit

D Additional Details for RQ2: Representation and Prediction Cost
- D.1 Additional Tokenization Granularity
- D.2 Tokenizer Provenance and Lineage
- D.3 Lineage Diagnostics
- D.4 Behavioral Evidence for Tokenizer Reuse
- D.5 Counterfactual Sentence Construction
- D.6 Prediction-Cost Evaluation
- D.7 Model Scope
- D.8 Semantic-Equivalence Control
- D.9 Base and Post-Trained Checkpoint Comparison

E Additional Details for RQ3: Generation
- E.1 Datasets and Sampling
- E.2 Prompting and Decoding
- E.3 Scoring and Outcome Measures
- E.4 Word-Length Adherence
- E.5 Response-Length Robustness

F Discussion and Practical Recommendations

G Extended Related Work

H Brief Historical Context

I Scope, Limitations, and Future Directions

A Regional Variant Resource

A.1 Construction and Filtering

We construct a curated resource of 1,813 matched AmE–Ref(BrE) variants for consistent measurement across the three research questions [RQ1 (Section 3), RQ2 (Section 4), and RQ3 (Section 5)]. The resource contains 1,538 orthographic pairs and 275 vocabulary pairs.

Variant pairs were manually compiled from linguistic references and web-based lexicons (Table 17). We merge entries across sources, normalize their representation, and remove duplicates before applying a strict matching criterion.

For analyses requiring direct word-level comparison, we retain one-to-one mappings in which each AmE form corresponds to one Ref(BrE) form. This criterion gives the resource a common unit of analysis across corpus frequency counts, tokenizer segmentation, and paired prediction-cost experiments. Multi-word and structurally different constructions remain important for regional variation, but are handled separately by DiAlign, whose local n-gram analysis is not restricted to the one-to-one lexicon.

Table 5: Distribution of the 1,813 matched AmE (🇺🇸) and Ref(BrE) (🇬🇧) variant pairs across linguistic categories in the curated resource. We report each category's share of the full inventory with representative examples, grouped into orthographic (*spelling*) and vocabulary-based contrasts.

Category	Difference Type	% of Pairs	Examples	
Orthographic/ Spelling	ends in "-or" (AmE) **vs.** "-our" (BrE)	2.26%	color (🇺🇸) **vs.** colour (🇬🇧)	labor (🇺🇸) **vs.** labour (🇬🇧)
	ends in "-ize" (AmE) **vs.** "-ise" (BrE)	11.58%	organize (🇺🇸) **vs.** organise (🇬🇧)	realize (🇺🇸) **vs.** realise (🇬🇧)
	ends in "-er" (AmE) **vs.** "-re" (BrE)	1.65%	center (🇺🇸) **vs.** centre (🇬🇧)	liter (🇺🇸) **vs.** litre (🇬🇧)
	ends in "-og" (AmE) **vs.** "-ogue" (BrE)	0.55%	dialog (🇺🇸) **vs.** dialogue (🇬🇧)	catalog (🇺🇸) **vs.** catalogue (🇬🇧)
	ends in "-ense" (AmE) **vs.** "-ence" (BrE)	0.22%	defense (🇺🇸) **vs.** defence (🇬🇧)	pretense (🇺🇸) **vs.** pretence (🇬🇧)
	"e" (AmE) **vs.** "ae" (BrE)	4.03%	esthetic (🇺🇸) **vs.** aesthetic (🇬🇧)	pediatric (🇺🇸) **vs.** paediatric (🇬🇧)
	words with single "l" **vs.** double "l"	8.88%	traveler (🇺🇸) **vs.** traveller (🇬🇧)	enroll (🇺🇸) **vs.** enrol (🇬🇧)
	sublexical spelling variation	49.75%	jewelry (🇺🇸) **vs.** jewellery (🇬🇧)	program (🇺🇸) **vs.** programme (🇬🇧)
Vocabulary	different lexical items entirely	21.07%	elevator (🇺🇸) **vs.** lift (🇬🇧)	flashlight (🇺🇸) **vs.** torch (🇬🇧)

A.2 Variant Grouping and Typology

To characterize the inventory, we assign every pair to one of ten mutually exclusive groups using a deterministic precedence order. These groups are then summarized under orthographic, vocabulary, and residual categories.

Group Definitions.

- **Group 1 (`-or` vs. `-our`):** suffix `-or` (AmE) ↔ suffix `-our` (Ref(BrE)).
- **Group 2 (`-ize` vs. `-ise`):** suffix `-ize` (AmE) ↔ suffix `-ise` (Ref(BrE)).
- **Group 3 (`-er` vs. `-re`):** suffix `-er` (AmE) ↔ suffix `-re` (Ref(BrE)).
- **Group 4 (`-og` vs. `-ogue`):** suffix `-og` (AmE) ↔ suffix `-ogue` (Ref(BrE)).
- **Group 5 (single `l` vs. double `ll`):** AmE single `l` ↔ Ref(BrE) double `ll`.
- **Group 6 (`-ense` vs. `-ence`):** suffix `-ense` (AmE) ↔ suffix `-ence` (Ref(BrE)).
- **Group 7 (`e` vs. `ae`):** the Ref(BrE) form contains `ae` where the AmE form uses `e`.
- **Group 8 (small orthographic edit):** equal-length forms with Levenshtein distance 1–2.
- **Group 9 (different lexical items):** forms with substantially different surface realization, identified by unequal length or edit distance greater than 2.
- **Group 10 (miscellaneous):** remaining pairs not captured by Groups 1–9.

Category Assignment. Groups 1–8 form the *Orthographic/Spelling* category, Group 9 forms the *Vocabulary* category, and Group 10 contains residual cases. This mutually exclusive grouping supports both aggregate comparisons and finer-grained analyses of recurring orthographic patterns.

Table 6: Representative distinctions between American English (AmE) (🇺🇸) and the British English reference, Ref(BrE) (🇬🇧), across orthography, vocabulary, grammar, and writing conventions. Examples reflect common majority-preference usage and are illustrative rather than exhaustive. Additional examples are provided in Tables 18 and 19 of the Appendix.

Level	American English (AmE) (🇺🇸)	Reference: British English (BrE) (🇬🇧)
Orthography	*color, center, organize, traveling, favorite, check, jewelry, program, catalog, defense*	*colour, centre, organise, travelling, favourite, cheque, jewellery, programme, catalogue, defence*
Vocabulary	*cell phone, crosswalk, downtown, train station, parking lot, ZIP code, vacation, apartment, elevator, line, sidewalk, flashlight, soda, popsicle*	*mobile phone, zebra crossing, city centre, railway station, car park, postcode, holiday, flat, lift, queue, pavement, torch, fizzy drink, ice lolly*
Grammar/usage	*I just ate*; *on the weekend*; *Do you have a pen?*; *The team is winning*; *different than*	*I've just eaten*; *at the weekend*; *Have you got a pen?*; *The team are winning*; *different from*
Conventions	double quotes; punctuation inside quotes; *December 31, 2024*; *first floor*; *11:15 PM*; $68°F$	single quotes; punctuation outside quotes; *31 December 2024*; *ground floor*; *11.15 pm*; $20°C$

A.3 Sources and Extended Examples

Table 6 summarizes representative AmE–Ref(BrE) contrasts across orthography, vocabulary, grammar, and writing conventions. Tables 18 and 19 provide a more detailed set of examples, while Table 17 lists the linguistic references and web-based resources used to construct the curated variant resource.

B DiAlign: Implementation, Validation, and Generalization

B.1 Implementation Details

This section records the implementation choices used for the DiAlign experiments in the paper.

Parameterization. We use the following configuration:

- n**-gram range:** $n \in \{2, 3, 4, 5\}$. This range captures multi-word and structural evidence while remaining within the maximum order supported by Google Books Ngram.
- **Temporal range:** $[y_{\min}, y_{\max}] = [1950, 2022]$. The broader window provides stable frequency estimates across low-frequency constructions. A contemporary-window analysis is reported in Section B.5.
- **Smoothing:** 0. We use the normalized yearly frequencies returned by the reference corpus without additional smoothing.
- **Lexicon boost:** $\beta = 1.5$, applied when a retained span contains an entry from the curated regional variant lexicon.

Frequency Lookup. For every retained span, we query the American English corpus (ID 17) and British English corpus (ID 6) in Google Books Ngram using the `requests` library (Lin et al., 2012). The service returns normalized yearly frequencies relative to the size of each corpus for that year. We average these normalized values over the selected temporal range.

Queries are case-insensitive and use exponential-backoff retries for transient failures. A persistent cache stores every queried $(n\text{-gram}, \text{corpus})$ pair, so repeated spans reuse previously retrieved frequencies. This substantially reduces reference-side queries when scoring large collections.

Filtering. We remove spans that are likely to reflect topic or named entities rather than regional usage. A candidate is discarded if it contains a person, organization, or location identified through NLTK named-entity chunking, or if every token in the span is a stopword.[5]

The resulting evidence consists of local constructions whose relative frequencies differ across the two reference distributions. These local signals are then combined with the divergence weighting and lexicon boost defined in Section 2.1.

[5] https://www.nltk.org

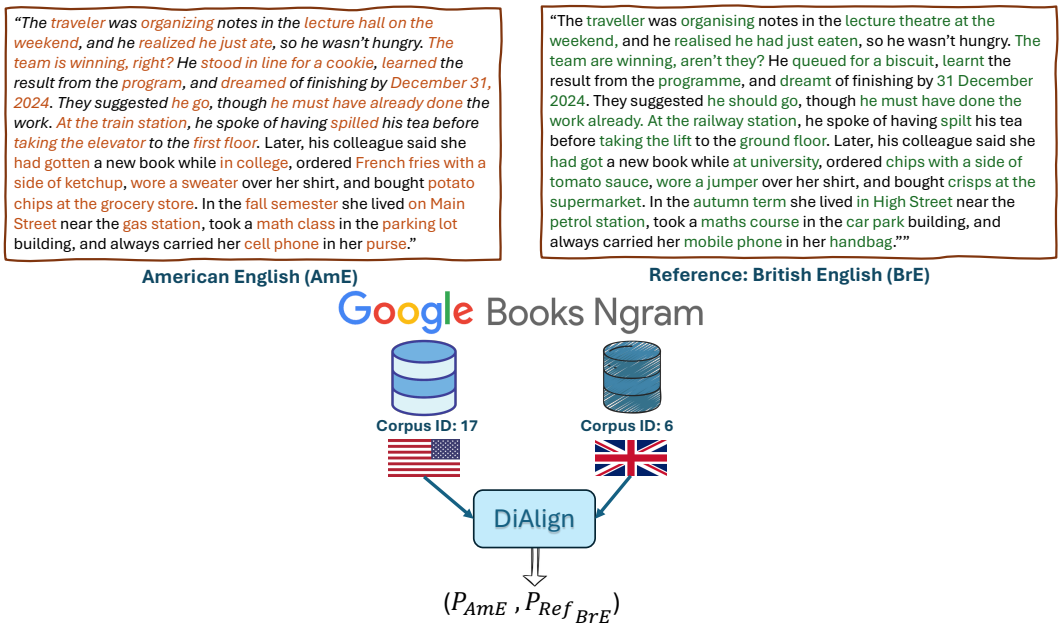


Figure 6: Illustrative walkthrough of DIALIGN. Parallel AmE and Ref(BrE) passages contain matched contrasts in spelling, vocabulary, grammar, conventions, and multi-word usage. DIALIGN decomposes each passage into overlapping 2–5-gram candidate spans, filters non-diagnostic spans, and compares the retained constructions against the American English (ID 17) and British English (ID 6) Google Books reference corpora. The resulting local evidence is aggregated into the regional alignment scores $(P_{\mathrm{AmE}}, P_{\mathrm{Ref}_{\mathrm{BrE}}})$.

B.2 ILLUSTRATIVE WALKTHROUGH

Figure 6 illustrates DIALIGN using parallel AmE and Ref(BrE) versions of the same passage. The example intentionally contains several types of regional variation so that the span-level behavior of the method is visible.

The passage includes examples such as:

- **Spelling:** *traveler/traveller*, *organizing/organising*, *realized/realised*, *program/programme*.
- **Vocabulary:** *cookie/biscuit*, *elevator/lift*, *train station/railway station*.
- **Verb morphology:** *dreamed/dreamt*, *learned/learnt*, *gotten/got*.
- **Tense and aspect:** *I just ate / I've just eaten*.
- **Collective noun agreement:** *The team is winning / The team are winning*.
- **Prepositional usage:** *on the weekend / at the weekend*.
- **Date conventions:** *December 31, 2024 / 31 December 2024*.
- **Institutional and everyday usage:** *in college / at university*, *cell phone / mobile phone*, *parking lot / car park*.

Because the passage is decomposed into overlapping 2–5-grams, these signals contribute locally rather than through whole-document similarity. The example also illustrates why DIALIGN can capture grammatical and collocational differences that are absent from the one-to-one variant inventory.

B.3 META-EVALUATION BENCHMARK

We evaluate DIALIGN on a balanced set of 1,500 short news passages with strong source provenance: 750 AmE samples and 750 Ref(BrE) samples.

Ref(BrE) Samples. We use English summaries from XL-Sum (Hasan et al., 2021), which is sourced from BBC News.[6] We select entries whose identifiers begin with `uk-england-` and randomly sample 750 summaries.

AmE Samples. We use the News Category Dataset (Misra, 2022), built from HuffPost articles. From the `U.S. NEWS` category, we randomly sample 750 `short_description` fields.[7]

Benchmark Statistics. The benchmark is therefore exactly balanced by reference variety. AmE passages contain 34.4 words on average and Ref(BrE) passages contain 24.7 words on average.

The labels are based on editorial provenance rather than separate sentence-level human annotation. BBC and HuffPost also differ in editorial style and topic distribution, so we treat the benchmark as a source-provenance validation set. Importantly, DiAlign does not score passages as whole documents: it decomposes each text into overlapping local 2–5-gram spans, removes named-entity and stopword-only spans, and aggregates the retained contrastive evidence. This design limits reliance on document-level topic, source identity, and global stylistic differences, while the benchmark tests whether local linguistic patterns track the expected regional convention.

B.4 Validation and Ablation

On the balanced benchmark, DiAlign reaches **93.18% accuracy**, **90.67% precision**, **96.25% recall**, and **93.38 F1** (Figure 2). Correct predictions are concentrated at larger confidence margins, while errors occur primarily near the decision boundary.

Table 7: Ablation study of DiAlign on the balanced AmE–Ref(BrE) meta-evaluation benchmark. We report classification performance (Accuracy, Precision, Recall, and F1) and mean confidence for AmE and Ref(BrE) predictions. Removing either divergence weighting (DW) or the boosting factor (BF) reduces performance, with the largest drop when both components are removed.

Ablations	Accuracy	Precision	Recall	F1 Score	Avg. Conf. (AmE)	Avg. Conf. (BrE)
DiAlign (final)	93.18	90.67	96.25	93.38	0.84	0.91
– w/o Divergence Weight (DW)	92.25	89.29	95.98	92.52	0.77	0.85
– w/o Boosting Factor (BF)	92.25	89.29	95.98	92.52	0.82	0.89
– w/o Both (DW + BF)	91.31	88.13	95.45	91.65	0.75	0.83

The ablation results in Table 7 show that both components contribute to the final estimator. Removing either divergence weighting or lexicon boosting reduces accuracy from 93.18% to 92.25%; removing both lowers it to 91.31%. The two components therefore provide complementary evidence rather than relying on a single lexical heuristic.

B.5 Temporal Robustness

Google Books contains text spanning many decades, so we test whether the alignment signal depends strongly on older usage. We rerun the complete meta-evaluation using only frequencies from `2000-2022`, while keeping all other scoring choices fixed.

The contemporary window produces nearly identical results: accuracy changes from 93.18% to 92.91%, and F1 from 93.38 to 93.12. The regional signal used by DiAlign is therefore stable across these two reference periods.

B.6 Generalizing DiAlign

DiAlign requires a pair of comparable reference distributions: one for the variety being studied and one for the reference variety. Google Books provides this pair directly for the AmE–Ref(BrE) analysis in this paper, but the scoring procedure itself is corpus-independent.

[6] https://huggingface.co/datasets/csebuetnlp/xlsum

[7] https://huggingface.co/datasets/heegyu/news-category-dataset

Table 8: Temporal robustness of DIALIGN to the Google Books reference window. Restricting the reference distributions from 1950–2022 to 2000–2022 changes accuracy by only 0.27 points and F1 by 0.26 points, indicating that performance is stable to the temporal range used for frequency estimation.

Reference period	Accuracy	Precision	Recall	F1
1950–2022	93.18	90.67	96.25	93.38
2000–2022	92.91	90.42	95.98	93.12

For example, an AmE–Indian English study could construct an Indian-English reference corpus from regional news organizations, institutional websites, and web sources with reliable provenance. Country-code top-level domains (ccTLDs), such as `.in` for India, can contribute one provenance signal when combined with publisher, institutional, and source-location metadata. Comparable U.S. sources could provide the AmE distribution. Both collections could then be converted into normalized 2–5-gram frequency tables and used directly by DIALIGN.

The same strategy applies to other regional varieties. Potential provenance signals include `.ng` for Nigeria, `.au` for Australia, `.nz` for New Zealand, and `.sg` for Singapore. Combining multiple signals is preferable to treating a domain suffix as a linguistic label on its own.

For any two varieties A and B, the core contrast becomes

$$\mathrm{LR}_{A,B}(g) = \log_2 \left(\frac{f_A(g)}{f_B(g)} \right).$$

Positive values provide evidence toward A, negative values toward B, and the remaining weighting and aggregation procedure is unchanged.

The framework can also be instantiated beyond English when suitable variety-specific corpora can be constructed. The central requirement is a pair of sufficiently large and comparable reference distributions with reliable provenance. Google Books is therefore the reference resource used for the present study, rather than a requirement of the method.

C ADDITIONAL DETAILS FOR RQ1: DATA EXPOSURE

C.1 PRETRAINING DATASETS

We audit six widely used pretraining corpora representing books, encyclopedic text, web data, and mixed-domain collections.

- **BookCorpus** (Zhu et al., 2015): a collection of unpublished novels frequently used in language-model pretraining. The analyzed release contains approximately 74 million documents and 1.28 billion tokens.[8]
- **Wikipedia** (Foundation, 2024): encyclopedic text from Wikipedia dumps. The analyzed release contains approximately 6.4 million documents and 4.3 billion tokens.[9]
- **Common Crawl (C4)** (Raffel et al., 2020): a cleaned and deduplicated Common Crawl collection containing approximately 365 million documents and 156 billion tokens.[10]
- **Falcon RefinedWeb** (Penedo et al., 2023): a Common Crawl-derived web corpus with additional filtering and deduplication, comprising approximately 968 million documents and 600 billion tokens.[11]
- **RedPajama** (Weber et al., 2024): a reproduction of major LLaMA-style data sources spanning books, code, papers, and web content. We use the sampled 1T release analyzed in this study.[12]

[8] https://huggingface.co/datasets/bookcorpus/bookcorpus
[9] https://huggingface.co/datasets/wikimedia/wikipedia
[10] https://huggingface.co/datasets/allenai/c4
[11] https://huggingface.co/datasets/tiiuae/falcon-refinedweb
[12] https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T-Sample

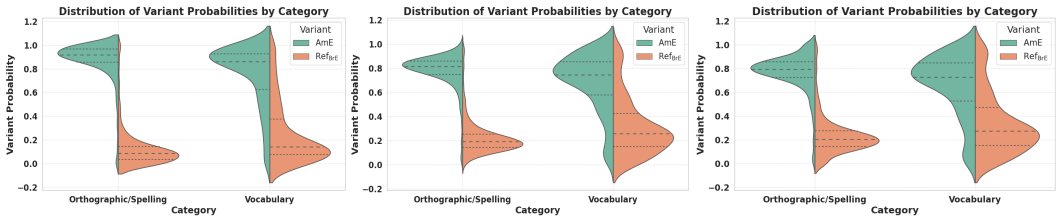


Figure 7: Variant-probability distributions for **(a)** BookCorpus, **(b)** Falcon RefinedWeb, and **(c)** RedPajama. Orthographic contrasts are more strongly concentrated toward AmE, while vocabulary contrasts are more dispersed but still lean toward AmE.

- **Dolma** (Soldaini et al., 2024): an open mixed-domain collection combining web pages, books, code, papers, forums, and social-media data. We use the v1.6 sample release.[13]

C.2 Pretraining Preprocessing

All six corpora are processed through the same normalization pipeline before variant counting. Text is lowercased, and HTML tags, hyperlinks, and email addresses are removed. Hyphens and slashes are replaced with whitespace so that joined forms do not obscure word boundaries. Non-alphabetic characters are then replaced with whitespace and repeated whitespace is collapsed.

This normalization yields a consistent representation for matching the 1,813 AmE–Ref(BrE) pairs across corpora. The same preprocessing is applied to both members of every pair, so corpus comparisons use an identical matching procedure.

C.3 Corpus Processing and Sampling

We process the large pretraining corpora using Hugging Face streaming, which allows examples to be analyzed without locally materializing the full dataset.[14]

Variant-frequency counts are computed over the streamed corpus material. For the broader DiAlign analysis, reference lookup is the dominant computational cost because every distinct local span requires corresponding AmE and Ref(BrE) reference frequencies. We therefore apply DiAlign to a uniform sample of 10,000 documents from each pretraining corpus. The same sampling size and scoring procedure are used across all six resources.

C.4 Additional Pretraining Results

Figures 7–9 provide the distributional views omitted from the main paper.

C.5 Post-Training Dataset Audit

Table 9 reports the dataset-level results underlying Table 2 in the main paper. The audit covers 21 public resources spanning supervised fine-tuning, chat data, human feedback, preference learning, reward modeling, and safety alignment.

Where multiple closely related releases exist, we use one cleaned, current, or original public release to avoid duplicate counting across dataset families.

For each dataset, **Weighted AmE (%)** reports the aggregate AmE share across all observed variant occurrences. **Orth. AmE (%)** and **Vocab. AmE (%)** instead report mean pair-level probabilities within their respective linguistic categories, so these columns are not complements of one another.

[13] https://huggingface.co/datasets/allenai/dolma
[14] https://huggingface.co/docs/datasets/en/stream

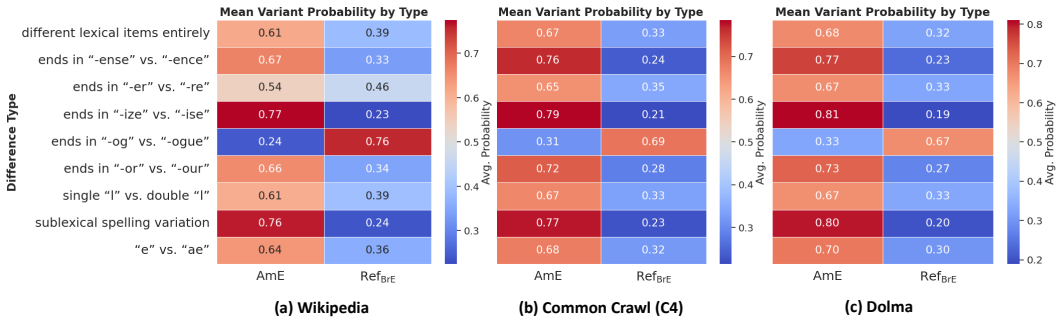


Figure 8: Category-level AmE and Ref(BrE) probabilities for **(a)** Wikipedia, **(b)** Common Crawl (C4), and **(c)** Dolma. Cells show the mean pair-level probability within each linguistic contrast category.

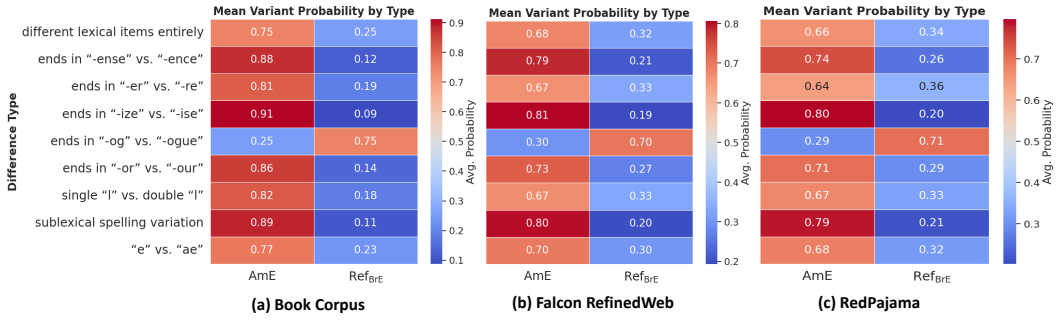


Figure 9: Category-level AmE and Ref(BrE) probabilities for **(a)** BookCorpus, **(b)** Falcon RefinedWeb, and **(c)** RedPajama. The same category-level pattern observed in Figure 8 is also evident across these corpora.

For OpenAI Summarize Comparisons, we remove the context-sensitive *mail/post* pair because *post* frequently denotes an online post rather than postal mail. After excluding this pair, the dataset contains 176,589 retained variant occurrences and has a weighted AmE share of 73.18%.

Across all 21 datasets, weighted AmE prevalence ranges from 65.29% in Databricks Dolly 15k to 89.72% in WizardLM Evol-Instruct v2.

D Additional Details for RQ2: Representation & Prediction Cost

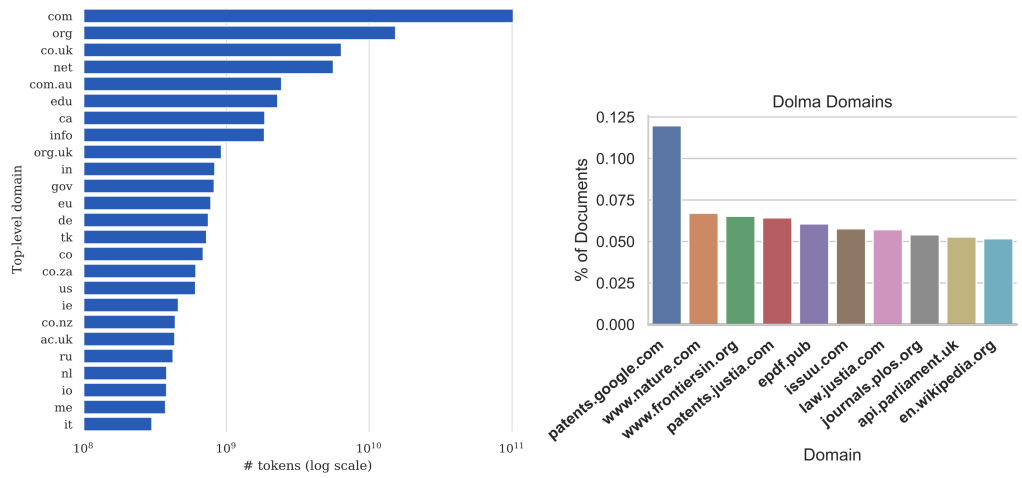
D.1 Additional Tokenization Granularity

Figure 11 extends the main granularity analysis to the remaining tokenizer families. The same long-tail pattern is visible across the panel: Ref(BrE) variants are more frequently split into three or more subword units.

D.2 Tokenizer Provenance and Lineage

The tokenizer results are best interpreted with vocabulary provenance separated from model developer origin. We therefore audit three related but distinct properties: model origin, tokenizer vocabulary lineage, and observed segmentation behavior.

Seven of the nine evaluated tokenizers use independent or source vocabularies. Llama-3.3 and StableLM-2 instead inherit or extend US-origin tokenizer artifacts. These tokenizers remain part of the behavioral analysis because their segmentation behavior is directly relevant to users and models, while their lineage provides important context for interpreting geographic metadata.



(a) Common Crawl (C4.EN) **(b) Dolma**

Figure 10: Source distributions in two pretraining corpora. **(a)** Top-level domains in C4.EN are concentrated in `.com` and `.org`, with much smaller shares from country-specific domains such as `.co.uk`. **(b)** High-frequency sources in Dolma illustrate the concentrated provenance of a curated mixed-domain corpus (Raffel et al., 2020; Soldaini et al., 2024).

Table 9: **Dataset-level post-training audit. Hits** counts occurrences matching either member of the curated AmE–Ref(BrE) variant pairs. **Weighted AmE (%)** reports the aggregate AmE share across observed hits, while **Orth. AmE (%)** and **Vocab. AmE (%)** report mean pair-level AmE shares within orthographic and vocabulary categories, respectively. **HF** links identify the exact Hugging Face dataset releases used in the audit.

Dataset	Stage	Rows	Hits	Weighted AmE (%)	Orth. AmE (%)	Vocab. AmE (%)	HF
Alpaca Cleaned (Taori et al., 2023)	SFT	51,760	74,991	85.35	94.84	80.05	HF
Databricks Dolly 15k (Conover et al., 2023)	SFT	15,011	19,997	65.29	78.09	67.65	HF
OpenAssistant Conversations v2 (Köpf et al., 2023)	SFT	61,278	61,030	83.08	92.56	76.85	HF
UltraChat 200k (Ding et al., 2023)	SFT	207,865	2,442,647	84.23	90.32	75.68	HF
OpenHermes 2.5 (Teknium, 2023)	SFT	1,001,551	2,517,105	81.81	85.78	74.01	HF
WizardLM Evol-Instruct v2 (Xu et al., 2024)	SFT	143,000	634,919	89.72	95.20	80.33	HF
No Robots (Rajani et al., 2023)	SFT	9,500	20,951	76.51	90.18	77.66	HF
LIMA (Zhou et al., 2023)	SFT	1,030	5,348	78.78	89.05	74.62	HF
FLAN Collection (Longpre et al., 2023)	SFT	1,386,050	2,099,656	69.78	76.20	69.85	HF
Super-NaturalInstructions (Wang et al., 2022)	SFT	3,082,094	4,479,160	72.71	80.41	73.01	HF
ShareGPT Vicuna (Chiang et al., 2023)	SFT	120,675	1,204,637	84.35	91.40	79.69	HF
OpenOrca (Mukherjee et al., 2023)	SFT	4,233,923	10,632,811	71.79	80.65	72.21	HF
Anthropic HH-RLHF (Bai et al., 2022)	RLHF	160,800	488,280	78.87	93.85	80.91	HF
OpenAI WebGPT Comparisons (Nakano et al., 2022)	RLHF	19,578	55,193	79.50	91.58	79.97	HF
Stanford Human Preferences (Ethayarajh et al., 2022)	RLHF	348,718	1,013,154	72.40	84.36	73.05	HF
UltraFeedback Binarized (Cui et al., 2024)	RLHF	61,135	390,601	82.69	90.46	77.77	HF
OpenAI Summarize Comparisons (Stiennon et al., 2020)	RLHF	92,534	176,589	73.18	85.31	74.52	HF
HelpSteer2 (Wang et al., 2024)	RLHF	20,324	77,332	84.72	92.87	75.87	HF
Nectar (Zhu et al., 2024)	RLHF	182,954	2,044,872	85.09	93.15	79.39	HF
PKU-SafeRLHF (Ji et al., 2025)	RLHF	73,907	178,469	86.86	90.76	77.96	HF
BeaverTails (Ji et al., 2023)	RLHF	300,567	242,377	83.48	92.52	79.63	HF
Overall	–	**11,574,254**	**28,860,119**	**76.17**	**88.55**	**76.22**	–

D.3 Lineage Diagnostics

Vocabulary overlap alone cannot distinguish direct inheritance from ordinary surface similarity. We therefore use three complementary diagnostics.

For tokenizer vocabularies V_i and V_j, directional vocabulary coverage is

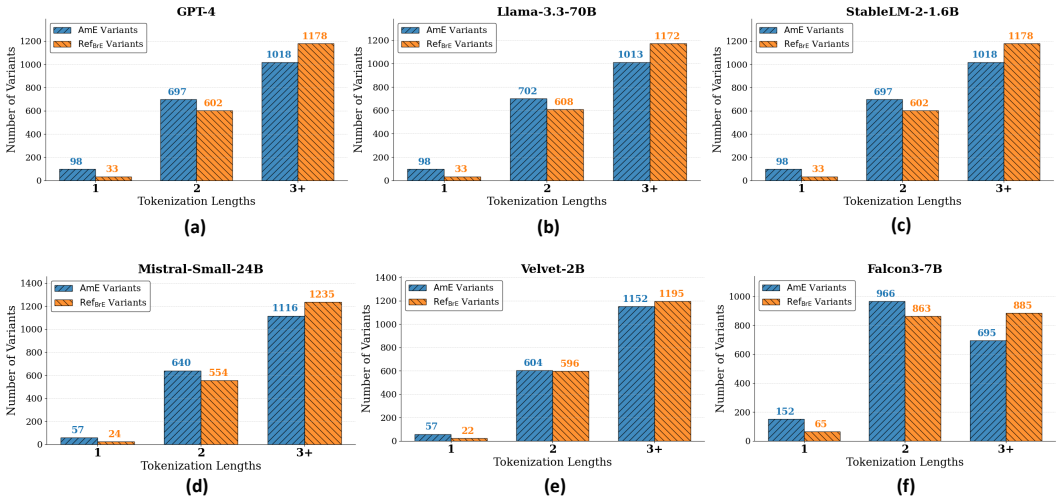


Figure 11: Additional tokenization-granularity results for **(a)** GPT-4, **(b)** Llama-3.3-70B, **(c)** StableLM-2-1.6B, **(d)** Mistral-Small-24B, **(e)** Velvet-2B, and **(f)** Falcon3-7B. Bars show the number of AmE and Ref(BrE) variants represented by 1, 2, or 3+ subwords.

Table 10: **Tokenizer provenance audit.** Model developer country and tokenizer provenance are reported separately. **Independent** refers specifically to tokenizer-vocabulary provenance, not to model weights, architecture, or training data. Derived tokenizers are retained in the behavioral analysis but are not counted as independent country-level evidence.

Tokenizer	Developer Country	Tokenizer lineage	Independent	Interpretation
GPT-4	USA (🇺🇸)	OpenAI `cl100k_base`	Yes	Independent US baseline
GPT-4o	USA (🇺🇸)	OpenAI `o200k_base`	Yes	Independent US baseline
Llama-3.3-70B	USA (🇺🇸)	Extended from US-origin tiktoken-style vocabulary	No	Derived; excluded from independent country-level evidence
Gemma-3-27B	USA (🇺🇸)	Gemma/Gemini-family vocabulary	Yes	Independent US tokenizer
DeepSeek-V3	China (🇨🇳)	DeepSeek BPE	Yes	Independent non-US tokenizer
Mistral-Small-24B	France (🇫🇷)	Mistral Tekken family	Yes	Independent non-US tokenizer
StableLM-2-1.6B	UK (🇬🇧)	Arcade100k/cl100k-like tiktoken BPE	No	Derived; not independent UK evidence
Velvet-2B	Italy (🇮🇹)	Velvet-family vocabulary	Yes	Independent non-US tokenizer
Falcon3-7B	UAE (🇦🇪)	Falcon3 BPE	Yes	Independent non-US tokenizer

$$S_{i \to j} = 100 \times \frac{|V_i \cap V_j|}{|V_i|}. \tag{2}$$

This quantity captures asymmetric extension, where one tokenizer may contain nearly the complete vocabulary of another while adding new tokens.

For shared token surfaces, exact rank/id agreement is

$$R_{ij} = 100 \times \frac{|\{v \in V_i \cap V_j : \mathrm{rank}_i(v) = \mathrm{rank}_j(v)\}|}{|V_i \cap V_j|}. \tag{3}$$

Finally, for inputs $x_1, \ldots, x_N$, token-boundary identity is

$$B_{ij} = 100 \times \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}\left\{\mathrm{boundaries}_i(x_k) = \mathrm{boundaries}_j(x_k)\right\}. \tag{4}$$

We compute boundary identity on the 1,813 AmE and 1,813 Ref(BrE) forms, as well as validation and generated-text inputs. The combination of vocabulary coverage, rank agreement, and actual segmentation behavior distinguishes artifact inheritance from superficial vocabulary overlap.

Table 11: **Tokenizer behavioral similarity with GPT-4.** Values report the percentage of strings for which each tokenizer produces the same token-boundary sequence as GPT-4. StableLM-2 and Llama-3.3 show near-perfect agreement, in contrast to the substantially lower agreement observed for independently sourced tokenizer families.

Tokenizer compared with GPT-4	AmE	BrE	Validation	Generated	All
StableLM-2-1.6B	**100.00**	**100.00**	80.80	81.17	91.87
Llama-3.3-70B	99.17	99.06	99.87	99.83	**99.43**
GPT-4o	74.85	77.05	40.27	29.00	58.58
DeepSeek-V3	67.02	70.22	47.80	25.08	55.42
Mistral-Small-24B	65.47	66.91	26.47	10.00	46.11
Falcon3-7B	58.58	58.91	11.93	5.17	37.48
Velvet-2B	20.63	35.58	20.20	9.08	22.62

D.4 Behavioral Evidence for Tokenizer Reuse

StableLM-2 provides the clearest inheritance case. Its tokenizer exhibits **100.00% token-boundary identity** with GPT-4 on both the AmE and Ref(BrE) variant sets. In addition, all `100,256` shared tokens in the GPT-4-sized base vocabulary retain identical rank/id assignments.

Llama-3.3 shows similarly strong behavioral continuity, with 99.17% boundary identity on AmE variants and 99.06% on Ref(BrE) variants relative to GPT-4.

The independently sourced DeepSeek, Mistral, Falcon, and Velvet tokenizers show substantially lower boundary identity. GPT-4o is also informative: despite considerable token-surface overlap with GPT-4, its rank structure and segmentation behavior differ markedly. Surface vocabulary overlap alone is therefore insufficient to establish lineage.

These results show how tokenizer conventions can persist across model families through artifact reuse. Model developer origin and tokenizer provenance therefore describe different parts of the model-development stack, and we report them separately throughout the paper.

D.5 Counterfactual Sentence Construction

The prediction-cost analysis uses the same `1,813` matched regional variant pairs, comprising `1,538` orthographic pairs and `275` vocabulary pairs. Instantiating these variants in controlled contexts produces `8,533` matched sentence-pair rows. Within each pair, only the target AmE or Ref(BrE) span changes.

Most pairs use conservative minimal-mention templates:

```
The word is {variant}.
This sentence uses the word {variant}.
The document contains the term {variant}.
```

Orthographic pairs additionally use spelling-specific contexts:

```
The article used the spelling {variant}.
The preferred form in this document is {variant}.
```

These templates isolate the regional surface form while holding the surrounding sentence context fixed. The broader controlled RQ2 resource uses core-quality spelling pairs instantiated with either minimal-mention or orthographic-context templates. This broader slice is also used for the semantic-equivalence analysis in Section D.8.

The strict confirmatory prediction-cost results reported in Table 12 use the more conservative minimal-mention subset.

We additionally instantiate 18 manually reviewed lexical pairs in natural contexts. For example,

```
She rented a small {variant} near the station.
```

Table 12: **Prediction-cost asymmetry on the tokenizer-controlled RQ2 slice.** Rows contain orthographic minimal-mention counterfactuals for which the AmE and $\mathrm{Ref}_{\mathrm{BrE}}$ sentences receive identical token counts under the evaluated model. **Gap/token** denotes the $\mathrm{Ref}_{\mathrm{BrE}}$ minus AmE difference in per-token negative log-likelihood, with positive values indicating higher prediction cost for $\mathrm{Ref}_{\mathrm{BrE}}$. Confidence intervals are paired-bootstrap 95% intervals.

Family	Stage	Rows	Gap / token	95% CI	$\mathrm{Ref}_{\mathrm{BrE}}$ higher
Llama	Base	1,008	0.189	[0.185, 0.193]	100.0%
Llama	Post	1,008	0.264	[0.262, 0.267]	100.0%
StableLM	Base	2,350	0.210	[0.204, 0.214]	100.0%
StableLM	Post	2,350	0.400	[0.389, 0.411]	100.0%
Ministral	Base	1,068	0.213	[0.208, 0.218]	100.0%
Ministral	Post	1,068	0.234	[0.229, 0.239]	100.0%
Gemma	Base	1,448	0.226	[0.222, 0.230]	100.0%
Gemma	Post	1,448	0.340	[0.329, 0.351]	100.0%
Falcon3	Base	2,350	0.303	[0.295, 0.310]	100.0%
Falcon3	Post	2,350	0.420	[0.408, 0.430]	100.0%

instantiates *apartment/flat*. These examples provide a separate natural-context robustness subset.

D.6 Prediction-Cost Evaluation

For each checkpoint, we compute sentence negative log-likelihood and normalize by the model-specific token count:

$$\Delta^{\mathrm{loss}}_{m,i} = \frac{\mathrm{NLL}_m(x_i^{\mathrm{Ref}_{\mathrm{BrE}}})}{|\tau_m(x_i^{\mathrm{Ref}_{\mathrm{BrE}}})|} - \frac{\mathrm{NLL}_m(x_i^{\mathrm{AmE}})}{|\tau_m(x_i^{\mathrm{AmE}})|}.$$

Positive values indicate higher prediction cost for the Ref(BrE) counterfactual.

For the strict confirmatory analysis, we restrict to core-quality spelling pairs instantiated with minimal-mention templates. We report results on this clean subset and on tokenizer-controlled slices. The strictest sentence-level control retains only pairs with equal token counts:

$$|\tau_m(x_i^{\mathrm{Ref}_{\mathrm{BrE}}})| = |\tau_m(x_i^{\mathrm{AmE}})|.$$

This ensures that the reported per-token loss gap is not attributable to one counterfactual being represented by more sentence tokens than the other.

We also evaluate a target-level control in which the whitespace-prefixed AmE and Ref(BrE) forms receive identical token counts. This captures tokenizers that encode word-in-context forms using leading-space units.

Alongside the mean gap, we report the directional share, defined as the percentage of pairs for which $\Delta^{\mathrm{loss}}_{m,i} > 0$.

Every evaluated checkpoint assigns higher per-token loss to the Ref(BrE) counterfactual on the strict equal-sentence-token-count slice, and every paired bootstrap 95% confidence interval excludes zero. The full evaluation contains `85,150` paired-loss records across ten checkpoints, with no skipped models.

D.7 Model Scope

The prediction-cost panel contains five model families for which matched public base and post-trained checkpoints are available at comparable scale.

DeepSeek-V3 remains part of the tokenizer and provenance analyses, but a comparable small DeepSeek-V3 base/post-trained pair is not publicly available. Velvet-2B is likewise retained in

Table 13: Matched base and post-trained checkpoints used for the RQ2 prediction-cost analysis, spanning five model families and ten checkpoints.

Family	Stage	Hugging Face checkpoint
Gemma	Base	google/gemma-3-1b-pt
Gemma	Post-trained	google/gemma-3-1b-it
Ministral	Base	mistralai/Ministral-3-3B-Base-2512
Ministral	Post-trained	mistralai/Ministral-3-3B-Instruct-2512
Llama	Base	meta-llama/Llama-3.1-8B
Llama	Post-trained	meta-llama/Llama-3.1-8B-Instruct
StableLM	Base	stabilityai/stablelm-2-1_6b
StableLM	Post-trained	stabilityai/stablelm-2-zephyr-1_6b
Falcon3	Base	tiiuae/Falcon3-7B-Base
Falcon3	Post-trained	tiiuae/Falcon3-7B-Instruct

Table 14: **Layerwise semantic-equivalence control.** Values report model-level mean cosine similarities averaged across ten evaluated checkpoints. Matched AmE–$\mathrm{Ref}_{\mathrm{BrE}}$ counterfactuals are more similar than character-perturbed and unrelated controls across layers, with the clearest separation at the final layer.

Layer	Comparison	Models	Mean cosine
Embedding	AmE–$\mathrm{Ref}_{\mathrm{BrE}}$ counterfactual	10	0.902
Embedding	Character perturbation	10	0.824
Embedding	Unrelated control	10	0.504
Middle	AmE–$\mathrm{Ref}_{\mathrm{BrE}}$ counterfactual	10	1.000
Middle	Character perturbation	10	0.999
Middle	Unrelated control	10	0.997
Final	AmE–$\mathrm{Ref}_{\mathrm{BrE}}$ counterfactual	10	0.989
Final	Character perturbation	10	0.928
Final	Unrelated control	10	0.814

the tokenizer analysis, while its model family does not provide a separately released public base checkpoint for the matched comparison.

D.8 Semantic-Equivalence Control

The semantic-equivalence analysis uses the broader controlled spelling slice described in Section D.5, including both minimal-mention and orthographic-context templates. Its purpose is to verify that changing the regional form preserves the intended sentence meaning.

Let $h_{m,\ell}(x)$ denote the mean-pooled representation of sentence x at layer ℓ. For each matched pair, we compute

$$\mathrm{sim}_{m,\ell,i}^{\mathrm{AmE},\mathrm{Ref}_{\mathrm{BrE}}} = \cos\Big(h_{m,\ell}(x_i^{\mathrm{AmE}}), h_{m,\ell}(x_i^{\mathrm{Ref}_{\mathrm{BrE}}})\Big).$$

Matched regional counterfactuals are compared against two controls: a character-perturbed AmE sentence and an unrelated Ref(BrE) sentence.

At the final layer, matched counterfactuals reach a mean cosine similarity of 0.989, compared with 0.928 for character perturbations and 0.814 for unrelated controls. The matched pairs remain closer at embedding, middle, and final representation levels.

The semantic evaluation contains `108,000` similarity records across the ten checkpoints, with no skipped models.

Prompt template for RQ3: Regional Preference in Generation

```
Answer the following question in {language}. Write a single, coherent
paragraph in plain text, using descriptive and open-ended language. Avoid
bullet points, lists, or formatting. Your response must be exactly
{WORD_LIMIT} words long—no more, no fewer. Count your words carefully.

Question: {question}
```

Figure 13: Prompt used under the two language conditions. We set `WORD_LIMIT`$= 50$ and `{language}` $\in$ {`English`, `British English (en-GB)`}.

D.9 Base and Post-Trained Checkpoint Comparison

Table 12 provides matched comparisons between base and post-trained checkpoints for all five model families on the strict minimal-mention, equal-sentence-token-count slice. In every family, the Ref(BrE) prediction-cost gap is larger in the post-trained checkpoint.

This is a consistent checkpoint-level pattern across Llama, StableLM, Ministral, Gemma, and Falcon3. Because each pair differs in its post-training procedure and data, the comparison characterizes the observed checkpoints rather than isolating a single post-training intervention.

E Additional Details for RQ3: Generation

E.1 Datasets and Sampling

RQ3 evaluates whether generated responses align more strongly with AmE or Ref(BrE) under neutral English prompting and explicit British-English steering.

We use two open-domain QA datasets with different communicative registers:

- **Natural Questions (NQ)** (Kwiatkowski et al., 2019): real search questions paired with Wikipedia-derived answers, providing a comparatively formal and encyclopedic setting.[15]
- **ELI5** (Fan et al., 2019): community question answering from Reddit, providing a more conversational and explanatory register.[16]

We discard questions shorter than five words and examples whose reference answer contains fewer than 30 words. Questions containing any of the 3,626 individual forms in the AmE–Ref(BrE) inventory are also removed so the prompt itself does not supply an explicit regional lexical cue.

From the remaining pool, we uniformly sample 300 NQ questions and 300 ELI5 questions.

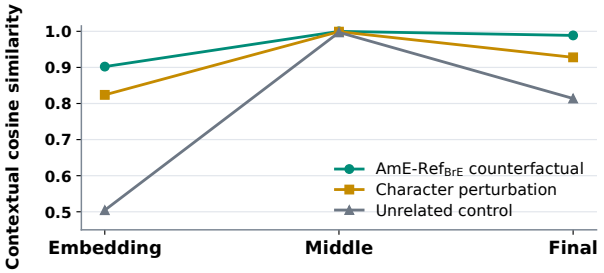


Figure 12: **Layerwise semantic-equivalence control.** Matched $\mathrm{AmE–Ref_{BrE}}$ counterfactuals remain closer than character-perturbed or unrelated controls across embedding, middle, and final representations, with the clearest separation at the final layer.

E.2 Prompting and Decoding

Each question is presented under two language conditions: `English` and `British English (en-GB)`. The requested output length is 50 words, providing a common target across models and conditions.

Decoding parameters are fixed across both conditions:

[15] https://huggingface.co/datasets/sentence-transformers/natural-questions
[16] https://huggingface.co/datasets/sentence-transformers/eli5

Table 15: Word-length adherence across models in RQ3 (Section 5). Each model was instructed to produce exactly 50 words per response. We report mean length, standard deviation, and range for Natural Questions (*formal*) and ELI5 (*informal*) under default English and British English (en-GB) prompting.

LLMs	Model Access	Natural Questions (NQ) [formal]				ELI5 [informal]			
		Default English (%AmE^{Default})		British English (%AmE^{BrE})		Default English (%AmE^{Default})		British English (%AmE^{BrE})	
		#Words Avg. [SD]	Range [min-max]	#Words Avg. [SD]	Range [min-max]	#Words Avg. [SD]	Range [min-max]	#Words Avg. [SD]	Range [min-max]
GPT-4o	🔒	50.19 [1.75]	[46–57]	50.32 [1.61]	[45–56]	50.41 [1.60]	[47–55]	50.35 [1.53]	[46–55]
Gemini-2.0-flash	🔒	52.74 [2.71]	[46–60]	51.92 [2.89]	[45–61]	53.10 [2.62]	[47–61]	52.82 [2.95]	[44–60]
Claude-3.7-sonnet	🔒	45.25 [2.29]	[39–52]	45.49 [2.35]	[39–57]	47.70 [2.45]	[42–56]	47.38 [2.32]	[41–54]
Llama-3.3-70B	🔓	41.68 [7.12]	[16–50]	41.47 [7.44]	[18–50]	41.33 [7.83]	[21–50]	40.10 [8.48]	[18–51]
Gemma-3-27B	🔓	49.03 [2.82]	[42–58]	49.19 [2.64]	[44–61]	49.18 [2.84]	[43–60]	49.90 [2.74]	[44–58]
DeepSeek-V3	🔓	53.03 [4.13]	[48–102]	53.40 [5.10]	[47–103]	53.60 [4.66]	[48–98]	53.33 [4.97]	[4–96]
Mistral-Small-24B	🔓	51.44 [18.80]	[13–318]	50.18 [10.97]	[20–88]	57.26 [11.68]	[23–120]	56.78 [10.43]	[34–101]
StableLM-2-1.6B	🔓	82.94 [39.90]	[8–364]	74.33 [40.15]	[8–351]	96.82 [24.23]	[23–199]	92.64 [24.76]	[41–199]
Velvet-2B	🔓	47.07 [30.94]	[8–406]	44.56 [23.03]	[8–135]	65.48 [18.55]	[22–122]	63.10 [17.25]	[31–120]
Falcon3-7B	🔓	44.01 [10.87]	[18–83]	41.72 [10.45]	[19–83]	49.41 [9.47]	[25–80]	48.14 [9.39]	[23–91]

$$\texttt{temperature} = 0.0, \qquad \texttt{top_p} = 1.0, \qquad \texttt{max_tokens} = 512.$$

Greedy decoding removes sampling variance, while the common decoding configuration ensures that the two language conditions differ only in the regional instruction.

E.3 Scoring and Outcome Measures

Each generated paragraph is scored using the same DiAlign procedure described in Section 2.1. For each response, we record P_{AmE}, $P_{\mathrm{Ref_{BrE}}}$, and the resulting majority alignment.

A response with no surviving directional evidence is treated as a zero-signal case and omitted from the aggregate classification rate. Such cases are rare and correspond to passages for which the retained local spans provide no measurable contrast between the two reference distributions.

For each model, dataset, and prompt condition, we report:

1. the percentage of responses classified as AmE, and
2. the mean AmE alignment score P_{AmE}.

E.4 Word-Length Adherence

The 50-word instruction is a common target rather than a post-processing constraint, so adherence varies by model. GPT-4o, Gemini, Claude, and Gemma remain relatively close to the requested length, while several smaller open-weight models exhibit wider distributions.

The resulting variation provides an additional opportunity to test whether response length itself explains the measured AmE preference.

E.5 Response-Length Robustness

DiAlign aggregates overlapping 2–5-gram evidence. For an input of $N \geq 5$ tokens, the number of candidate spans before filtering is

$$|\mathcal{G}(x)| = 4N - 10.$$

This gives short responses a comparatively dense set of overlapping local constructions while the fixed-length instruction keeps the intended generation scale comparable across models.

We additionally examine the outputs by realized word length:

Table 16: Realized response length and AmE alignment across length bins. Default-English outputs remain AmE-majority throughout, while British-English prompting produces the lowest AmE alignment among shorter, more instruction-adherent responses.

Condition	Length bin	Avg. words	DiAlign AmE (%)
Default	1–50	44.08	67.64
Default	51–75	55.82	67.33
Default	76–100	86.70	64.31
Default	101–150	116.03	62.54
Default	151+	225.60	61.01
British-English steer	1–50	43.62	47.12
British-English steer	51–75	55.99	48.44
British-English steer	76–100	86.43	58.80
British-English steer	101–150	116.26	57.79
British-English steer	151+	192.20	60.94

The aggregate association between response length and AmE alignment is very small: $r = 0.017$ across all outputs, $r = -0.043$ under default English prompting, and $r = 0.057$ under British-English steering.

Response length therefore does not explain the aggregate AmE preference. The length-bin analysis also shows that explicit British-English prompting has its strongest effect among shorter responses, while longer steered generations move somewhat closer to the AmE distribution.

F Discussion and Practical Recommendations

The AmE advantage appears consistently across data exposure, representation, and generation. RQ1 shows AmE overrepresentation in both pretraining and post-training resources, RQ2 identifies corresponding asymmetries in tokenization and model prediction cost, and RQ3 shows that AmE remains the dominant generation default under neutral English prompting. These findings suggest that regional imbalance should be addressed across the LLM development pipeline rather than only through output-level correction.

Audit and Rebalance Training Data. RQ1 (§3) shows that regional skew is already present in both pretraining and post-training data. Dataset construction can therefore track regional English variation alongside conventional measures of domain, quality, and language composition. For web-scale resources, provenance signals such as publisher location, institutional metadata, source URLs, and country-code top-level domains can support regional sampling strategies. For example, `.uk`, `.in`, `.ng`, `.au`, and `.sg` can provide useful regional signals when combined with source and publisher metadata rather than being treated as linguistic labels on their own.

The same audit should extend beyond pretraining. Instruction-tuning, preference, reward-model, and safety datasets can introduce additional regional skew after pretraining. Variant-frequency counts provide a lightweight way to monitor explicit spelling and vocabulary differences, while DiAlign can complement them with broader grammatical, structural, collocational, and multi-word evidence (§2.1). These diagnostics can be applied before training to compare candidate mixtures and adjust sampling when a strong unintended imbalance is present.

Include Regional Coverage in Tokenizer Design. RQ2 (§4) shows that Ref(BrE) forms are generally segmented less compactly than matched AmE forms. Tokenizer evaluation can therefore include regional coverage alongside aggregate compression metrics. Fertility and granularity provide simple diagnostics for identifying forms that are repeatedly split into longer subword sequences.

When such gaps are large, vocabulary extension or tokenizer retraining can increase coverage of underrepresented variants (Tejaswi et al., 2024). For example, forms that consistently fall into the 3+ subword category while their matched AmE counterparts are encoded more compactly are natural candidates for closer inspection. Tokenizer provenance should also be considered before reuse, since

our lineage analysis shows that segmentation behavior can persist across model families through inherited tokenizer artifacts.

Evaluate Prediction Cost Beyond Tokenization. More balanced segmentation does not necessarily imply equal model preference. Our controlled likelihood analysis shows that Ref(BrE) counterfactuals receive higher prediction cost even when sentence token count is identical. A practical regional evaluation can therefore report both tokenization differences and Δ^{loss}.

For example, a model may tokenize *color* and *colour* using the same number of sentence tokens while still assigning the Ref(BrE) counterfactual higher per-token negative log-likelihood. Evaluating both quantities helps distinguish tokenizer-level representation from preferences encoded in the model itself. Repeating the same evaluation for base and post-trained checkpoints can also reveal whether the regional gap becomes larger or smaller across model development.

Treat Regional Prompting as Steering. RQ3 shows that explicit British-English prompting generally moves outputs toward Ref(BrE), but the size of the shift varies across models and registers. Regional prompting is therefore useful as an interface control, but it should not replace upstream data and representation audits.

Evaluation should also go beyond simple spelling substitution. A response may use *colour* instead of *color* while still following AmE preferences in vocabulary, grammar, conventions, or multi-word usage. Regional settings should therefore be evaluated across several types of variation rather than judged only by a small set of visible lexical changes.

Use DIALIGN for Regional Auditing. DIALIGN can support these checks during both data construction and generation evaluation. During corpus building, it can estimate the regional alignment of candidate training data or synthetic additions. During evaluation, it can compare neutral-English and region-specific generations using the same local evidence.

The framework is also adaptable beyond the present AmE–Ref(BrE) comparison. For example, an AmE–Indian English analysis could construct a regional reference corpus using Indian news, institutional websites, and reliable provenance signals such as `.in`, combined with publisher and source-location metadata. Similar strategies could support studies of Nigerian, Australian, New Zealand, or Singapore English (Appendix B.6).

These recommendations map directly to the stages examined in this paper: audit regional exposure in training data, measure asymmetry in tokenization and model prediction cost, and evaluate the realized generation default. Study limitations and directions for extending the analysis are discussed in Appendix I.

G Extended Related Work

Training data audits and curation. The composition of training data has substantial effects on model behavior and generalization. The *Pretrainer's Guide* isolates the influence of data age, domain coverage, quality, and toxicity (Longpre et al., 2024), while corpus audits have exposed duplication, benchmark contamination, low-quality content, and gaps in provenance. *WIMBD*, for example, audits C4 and RedPajama for contamination and undesirable content (Elazar et al., 2024), and *Data Portraits* develops membership-testing tools for tracing training data (Marone & Durme, 2023). Broader analyses also document provenance imbalances across text, speech, and video resources, including heavy reliance on web-crawled and Western-centric sources (Longpre et al., 2025).

A parallel line of work studies how training corpora can be constructed and selected more effectively. This includes practical recipes for trillion-token datasets (Parmar et al., 2024), LLM-assisted quality selection through *QuRating* (Wettig et al., 2024), register-aware sampling (Myntti et al., 2025), domain-based organization of web text (Wettig et al., 2025), and reuse of filtered web content (Nguyen et al., 2025). Our analysis adds within-English variation to this auditing perspective by measuring AmE representation across pretraining and post-training resources and following the same directional asymmetry into later stages of the pipeline.

Tokenization and representational asymmetry. Tokenization determines how linguistic forms are mapped into the discrete units consumed by language models. Prior work documents large disparities in subword segmentation across languages and scripts (Nayeem et al., 2025), with consequences for sequence length, effective context capacity, latency, and monetary cost (Petrov et al., 2023; Alqahtani et al., 2026). Tokenization length and corpus frequency can also correlate with demographic properties of personal names, disproportionately fragmenting underrepresented forms (An & Rudinger, 2023). Related robustness studies show that lexical alternations, such as brand and generic drug names, can produce substantial variation in model behavior (Gallifant et al., 2024).

Other work examines the mechanisms and mitigation of such disparities. In machine translation, training distributions and subword segmentation can jointly disadvantage female and non-stereotypical gender inflections, while targeted token-embedding adaptation can reduce these gaps (Iluz et al., 2023). Controlled interventions further show that changing a word from a multi-token representation to a single token can substantially alter its probability (Lesci et al., 2025). Parity-Aware Byte-Pair Encoding similarly trades a small amount of compression for more balanced token counts across languages (Foroutan et al., 2025). Our analysis examines the same representational layer within English through tokenization efficiency, tokenizer provenance, and paired prediction cost.

English variation in NLP and LLMs. Research on language variation has documented systematic performance differences across regional and social varieties of English, particularly African American English (AAE) and South Asian Englishes (SAsE) (Demszky et al., 2021; Holt et al., 2024; Joshi et al., 2025). For AAE, disparities have been reported in part-of-speech tagging (Jørgensen et al., 2016), language identification (Blodgett et al., 2016), sentiment analysis (Kiritchenko & Mohammad, 2018), dependency parsing (Blodgett et al., 2018), hate-speech detection (Sap et al., 2019), and natural-language understanding (Ziems et al., 2022). Recent studies also show that LLMs can reproduce negative stereotypes associated with AAE (Hofmann et al., 2024; Fleisig et al., 2024).

User-centered studies report that SAsE speakers encounter lower reliability for their linguistic usage (Holt et al., 2024), while *Multi-VALUE* evaluates models across many English varieties and linguistic features, revealing broad robustness gaps (Ziems et al., 2023). Even differences between standardized English conventions affect model behavior: retrieval performance decreases when queries and documents use different spelling conventions (Chari et al., 2023), and language models respond differently to observed and novel spelling variants (Nielsen et al., 2023).

Our study places the American-English default within this literature on linguistic variation, but follows it across the development pipeline rather than a single downstream task. We examine its prevalence in pretraining and post-training data, its expression in tokenization and model prediction cost, and its realization in generations. This pipeline-level analysis provides a common empirical view of how the same AmE advantage recurs across data exposure, representation, and generation.

H Brief Historical Context

The divergence between American English (AmE) and British English (BrE) reflects both deliberate standardization and broader historical change. Samuel Johnson's 1755 dictionary helped codify spelling conventions that became influential in British literary and educational practice. In the United States, Noah Webster's 1828 dictionary promoted distinct and often simplified spellings as part of an emerging American linguistic identity (Baker, 2017). These processes established many of the orthographic contrasts still visible today.

The two standards subsequently acquired international reach through distinct geopolitical trajectories. British colonial expansion embedded BrE conventions in administration, education, journalism, and law across Africa, Asia, the Caribbean, and the Pacific. These conventions often persisted after political independence and remain institutionally influential across much of the Commonwealth and in English-using contexts such as India, Nigeria, Singapore, Australia, and New Zealand (Figure 14) (Calabrese et al., 2015; Tikly, 2016; Phillipson, 2018). This institutional footprint also persists through UK-linked language and educational institutions, including the British Council and international English assessment through IELTS.

American English expanded especially rapidly during the twentieth century through U.S. economic influence, mass media and global entertainment, technological development, and global



Figure 14: Timeline of independence across countries formerly under British colonization. The map highlights the wave of decolonization in the mid-twentieth century, when nations in Africa, Asia, the Caribbean, and the Pacific gained sovereignty. This geopolitical shift marked the decline of direct colonial governance but reinforced the institutional legacy of *British English (BrE)* in education, government, journalism, and law across many of these regions (Tikly, 2016; Phillipson, 2018).

commerce (Crystal, 2003; Gonçalves et al., 2018; Nordquist, 2024). Hollywood, popular culture, digital platforms, and technology ecosystems further strengthened its international visibility, particularly in contemporary online communication.

This history makes the AmE–BrE contrast useful for controlled analysis. Both are extensively documented standard varieties with systematic differences across spelling, vocabulary, grammar, conventions, and multi-word usage. BrE also provides a meaningful reference for institutional norms inherited by several other English varieties, while AmE has become especially prominent in the digital ecosystem from which modern LLMs are trained. Google Books provides separate reference corpora for the two varieties, allowing DiAlign to measure regional alignment beyond the curated one-to-one variant inventory.

I Scope, Limitations, and Future Directions

Our experiments use controlled comparisons designed to isolate the American-English advantage across multiple stages of the LLM pipeline. This design supports high-precision, cross-stage evidence and defines clear directions for broader generalization.

Reference Variety and Broader English Variation. We use Ref(BrE) as a controlled, globally meaningful comparator because it provides systematic contrasts with AmE across spelling, vocabulary, grammar, and usage that can be carried consistently from corpus statistics through tokenization, prediction cost, and generation. This prioritizes direct cross-stage comparability while providing a template for broader regional analysis. The same pipeline can be extended to Indian, Nigerian, Singapore, Australian, New Zealand, and other English varieties when suitable reference resources are available. Appendix B.6 describes how DiAlign can be instantiated with alternative corpora.

Lexicon Precision and Coverage. The 1,813-pair resource is designed for high-precision one-to-one comparison, which is especially important for tokenizer and counterfactual likelihood experiments where both regional forms must be directly comparable. We complement this resource with DiAlign, which captures grammatical, structural, collocational, stylistic, and multi-word evidence beyond the fixed lexicon. RQ1 uses type-level corpus counts for curated variants, with context-sensitive cases removed when they materially affect interpretation. Contextual modeling of larger web-scale variant inventories provides a direct extension.

Reference Corpora and Meta-evaluation. Google Books provides separate American and British English distributions, making it a consistent contrastive reference for DIALIGN. Normalized regional frequencies allow direct comparison of the same local constructions across the two reference distributions. DIALIGN decomposes each input into overlapping 2–5-gram spans and aggregates retained contrastive evidence at the local level. The resulting signal generalizes across domains: DIALIGN reaches 93.18% accuracy on the external BBC–HuffPost news benchmark, shows the same directional AmE pattern across encyclopedic, web, and mixed-domain pretraining corpora, and extends to model-generated open-domain responses in RQ3. Separately, the `2000–2022` robustness analysis yields nearly identical performance to the `1950–2022` setting, demonstrating temporal stability of the regional contrast. A matched-domain benchmark with independently annotated variety labels would further broaden evaluation across source conditions.

Generation Setting. RQ3 uses two contrasting open-domain registers under a fixed response-length protocol to isolate regional preference while controlling generation length. Natural Questions provides a more formal, encyclopedic setting, while ELI5 provides a more conversational one, and the observed preference is stable to realized response length (Section E.5). Long-form dialogue, specialized professional domains, mixed-variety inputs, and creative generation provide natural extensions for testing the default under more complex discourse and user contexts.

Tokenizer and Checkpoint Coverage. Our tokenizer panel spans multiple model families, vocabulary sizes, and development contexts, while the provenance audit separates independent tokenizers from inherited infrastructure. Prediction-cost experiments focus on five model families with matched public base and post-trained checkpoints, prioritizing within-family comparability over nominal model count. Larger closed models and families without comparable public checkpoints remain represented where artifact access permits, including tokenizer and generation experiments.

Attribution Across Pipeline Stages. The study traces where a recurring AmE asymmetry becomes visible across data exposure, tokenization, prediction cost, and generation. The consistency across these stages establishes a structural pattern. Causal decomposition requires controlled interventions such as retraining with rebalanced data, tokenizer replacement, or matched post-training interventions, providing a natural next step from structural tracing to causal attribution.

Table 17: Key linguistic and web-based sources used to construct the AmE–Ref(BrE) variant lexicon. Variant pairs were manually compiled across multiple sources, merged, and deduplicated to form a unified reference set for consistent analysis.

Source	Title (*linked*)	Description
Wikipedia	American and British English spelling differences	A widely cited reference outlining systematic orthographic differences between American and British English. The page provides examples of variant spellings (e.g., *color* vs. *colour*), historical background, and explanations of regional conventions. It served as one of the authentic linguistic resources for curating consistent one-to-one variant pairs in our lexicon.
ThoughtCo.	American English to British English Vocabulary	A curated reference list of American and British English vocabulary differences, created by experienced educators and subject experts. Provides reliable lexical contrasts in an accessible format, supporting the construction of our AmE–BrE lexicon.
Research Article	Mapping the Americanization of English in Space and Time	An empirical study tracing how American English variants spread globally across regions and over time. Offers quantitative evidence of AmE–BrE lexical contrasts, providing authoritative grounding for the curated variant pairs in our unified lexicon.
IELTS	British vs. American English in the IELTS Test: Key Differences	An official IELTS guide highlighting key vocabulary, spelling, and grammar differences between AmE and BrE. The resource systematically documents contrasts across domains such as food, school, homes, and grammar, making it a practical reference for understanding standardized English variations.
Grammarly	How to Select Your English Dialect	A practical guide from Grammarly explaining how to switch between English dialects in writing tools, highlighting spelling, vocabulary, and usage variations (AmE vs BrE). Because it enumerates common dialectal choices in real writing, it serves as a useful supplementary resource for identifying variant pairs.
SpellZone	Sixty American English Words and their British English Counterparts	SpellZone provides a practical reference list of 60 common AmE–BrE word pairs, illustrating clear lexical contrasts in spelling and vocabulary. The resource highlights straightforward one-to-one mappings useful for systematic dialectal analysis.
IELTS	Differences between British vs. American English	A guidance article from IELTS that outlines vocabulary, spelling, and grammatical contrasts between AmE and BrE, emphasizing how learners must maintain internal consistency between the dialects. This resource helps validate by showing differences accepted in international testing and educational settings.
SpellZone	Differences between British and American English spelling	It provides an overview of common orthographic contrasts (e.g. "colour/-color", "centre/center", "-re" vs "-er") between BrE and AmE. This resource was used as a web-based lexicon support to validate our curated variant pairs.
British Council	Differences between British and American English	An educational article by the British Council outlining vocabulary, grammar, and spelling distinctions between British and American English. It supports validation of variant pairs and highlights pedagogically recognized dialectal contrasts.
IELTS Liz	UK US Spelling Main Differences	A practical guide by IELTS Liz summarizing the core orthographic differences between British and American spelling. This resource helps cross-check variant consistency and supports the curated lexicon's alignment with real exam-related usage.
Word Finder	British vs. American English Words	A comparative list of British and American English words, highlighting more than just spelling shifts ("-u") and covering vocabulary contrasts in everyday usage. It offers additional variant candidates and informs our lexicon selection process.
Language Gallery	British VS American Spelling: What's the Difference?	A language-education blog article detailing common orthographic differences between British and American English (e.g., "realise/realize," "theatre/theater"). It served as a supplementary web-based lexicon to inform our manual variant curation.

Table 18: Representative AmE–Ref(BrE) distinctions across orthography, grammar, and formatting conventions. Entries reflect common majority-preference usage and are illustrative rather than exhaustive.

Category	British English (BrE)	American English (AmE)
o vs. ou	*colour, honour, behaviour*	*color, honor, behavior*
-re vs. -er endings	*centre, fibre, theatre*	*center, fiber, theater*
-ise vs. -ize endings	*recognise, authorise*	*recognize, authorize*
-yse vs. -yze endings	*analyse, paralyse, catalyse*	*analyze, paralyze, catalyze*
Single vs. double l (inflection)	*travelled, counselled*	*traveled, counseled*
-ll + -ly suffix	*skilfully, wilfully*	*skillfully, willfully*
Composite vowels	*anaesthetic, diarrhoea, paediatric, oestrogen*	*anesthetic, diarrhea, pediatric, estrogen*
Final silent -e/-ue	*catalogue, analogue, axe*	*catalog, analog, ax*
Silent -e before suffix	*ageing, likeable*	*aging, likable*
-ce vs. -se (noun/verb)	*licence (n), practise (v)*	*license (n/v), practice*
-ce vs. -se nouns	*defence, offence*	*defense, offense*
Programme vs. program	*TV programme, postgraduate programme*	*TV program, graduate program*
Orthographic pairs	*grey, cheque, manoeuvre, tyre, storey*	*gray, check, maneuver, tire, story (floor)*
Directional suffix -ward(s)	*towards, forwards, upwards*	*toward, forward, upward*
Sceptic/k alternation	*sceptic, sceptical*	*skeptic, skeptical*
Judgement spelling	*judgement*	*judgment*
Maths/Math	*maths*	*math*
Season name	*autumn*	*fall*
Present perfect vs. past	*I've just eaten.*	*I just ate.*
Mandative subjunctive	*They suggested he* should *apply.*	*They suggested he apply.*
shall vs. will	*I shall go tomorrow.*	*I will go tomorrow.*
Irregular verb morphology	*learnt, dreamt, spoilt*	*learned, dreamed, spoiled*
Collective noun agreement	*The team are winning.*	*The team is winning.*
Possession verb	*I've got a car.*	*I have a car.*
Got vs. gotten	*He's got very tired.*	*He's gotten very tired.*
Prepositional usage	*at the weekend*, *in a team*	*on the weekend*, *on a team*
Tag questions	*You're ready, aren't you?*	*You're ready, right?*
Subjunctive usage	*They suggested he should go.*	*They suggested he go.*
Auxiliary ellipsis	*He must have done.*	*He must have.*
Numerals ("and")	*one hundred and twenty*	*one hundred twenty*
Restrictive relative marker	*the report* which *was submitted*	*the report* that *was submitted*
Possession questions	*Have you got a pen?*	*Do you have a pen?*
Necessity negative	*You needn't attend.*	*You don't need to attend.*
Difference construction	*different from / different to*	*different from / different than*
Quotation marks	Prefers single quotes '...q'	Prefers double quotes "..."
Commas/periods in quotes	Outside the closing quotes	Inside the closing quotes
Abbreviations with periods	*Mr, Dr*	*Mr., Dr.*
Oxford/serial comma	Rare	Common
Date format (written)	*31 December 2024*	*December 31, 2024*
Date punctuation (written)	*19 September 1973*	*September 19, 1973*
Date format (numeric)	*31/12/2024 (DD/MM/YYYY)*	*12/31/2024 (MM/DD/YYYY)*
Legal/institutional terms	*Ministry of Defence*	*Department of Defense*
Institutional article usage	*in hospital; at university*	*in* the *hospital; at* the *university*
Floor numbering	*ground floor, first floor (one up)*	*first floor, second floor (one up)*
Time notation	*11.15 pm; 23.15 common*	*11:15 PM; 24-hour less common*

Table 19: Representative AmE–Ref(BrE) preferences across everyday domains, including transport, household, and food. Entries reflect common majority-preference usage and are illustrative rather than exhaustive.

Category	British English (BrE)	American English (AmE)
Preposition before days	*She resigned on Thursday.*	*She resigned Thursday.*
Street naming	*in the High Street*	*on Main Street*
Transitivity (protest)	*protest against discrimination*	*protest discrimination*
Ditransitives (write)	*write to me*	*write me*
Meeting collocation	*meet the team*	*meet with the team*
Transport & wayfinding		
Pedestrian crossing	*zebra crossing*	*crosswalk*
Junction type	*roundabout*	*traffic circle / rotary*
Road maintenance	*roadworks*	*road work*
Parking payment	*pay and display*	*metered parking*
Perimeter road	*ring road*	*beltway*
Vehicle hire	*hire car / car hire*	*rental car / car rental*
Estate car vs. wagon	*estate car*	*station wagon*
Household & services		
Postal addressing	*postcode*	*ZIP code*
Carry-on baggage	*hand luggage*	*carry-on*
Washing liquid	*washing-up liquid*	*dish soap*
Waste container	*dustbin*	*trash can / garbage can*
Clothes washer	*washing machine*	*washer*
Cash dispenser	*cashpoint*	*ATM*
Public convenience	*public toilet*	*restroom*
Mobile device	*mobile phone*	*cell phone*
Food & drink		
Confection	*candyfloss*	*cotton candy*
Frozen treat	*ice lolly*	*popsicle*
Leafy green	*rocket*	*arugula*
Soft drink	*fizzy drink*	*soda*
Allium term	*spring onion*	*green onion / scallion*
Cake term	*fairy cake*	*cupcake*
Places & urban terms		
City core	*city centre*	*downtown*
Real estate profession	*estate agent*	*realtor / real estate agent*
Holiday lodging	*holiday let*	*vacation rental*
Queueing term	*post office queue*	*post office line*
Queuing expression	*join the queue*	*wait in line*
Public transport info	*railway timetable*	*train schedule*
Education & work		
Practical training	*work placement*	*internship*
Assessment term	*marking scheme*	*grading rubric*
Student level	*first-year student*	*freshman*
Residence	*halls of residence*	*dorm / residence hall*
Single-word vocabulary		
flat / apartment	*flat*	*apartment*
lorry / truck	*lorry*	*truck*
pavement / sidewalk	*pavement*	*sidewalk*
wardrobe / closet	*wardrobe*	*closet*
lift / elevator	*lift*	*elevator*
petrol / gas	*petrol*	*gas*
railway / railroad	*railway*	*railroad*
holiday / vacation	*holiday*	*vacation*