The Dialect Tax: Dialectal Biases Persist throughout the Language Modeling Pipeline

Elle
University of Oxford, Department of Computer Science
elle.yang@cs.ox.ac.uk

Abstract

Systematic dialectal performance gaps in language models (LMs) are well documented, but the source of these disparities within the modern language modeling pipeline remains unclear. Our study traces this "dialect tax" across the natural language processing pipeline. Using parallel English dialect corpora that hold meaning fixed while varying surface form, we first confirm that LMs recognize matched Standard American English (SAE) and dialectal texts as semantically equivalent. However, we discover further representational gaps corresponding to downstream performance gaps. Across model families and generations, modern LMs still encode dialectal texts unequally during tokenization, pre-training, post-training, and inference. Strikingly, bypassing traditional subword segmentation via a character-level counterfactual tokenizer removes neither input and output asymmetries nor dialectal accuracy gaps. During pre-training, dialect pairs induce more divergent gradient updates than pairs of entirely unrelated SAE documents, indicating that models find semantically equivalent dialectal content harder to learn from than unrelated SAE documents. During post-training, reward models show contextual, unstable dialect preferences, assigning higher values to isolated AAVE-exclusive tokens than to SAE-exclusive tokens, while full reasoning contexts receive task- and model-dependent dialect penalties. Overall, our findings suggest that the dialect tax is encoded and accumulated not by any one step in isolation, but at every step of the language modeling process.

1 Introduction

AI systems, particularly language models (LMs), have become technological staples worldwide. Although advances in deep learning for natural language processing (NLP) have improved the performance of these systems, these technologies continue to exhibit societal biases at every stage of the language modeling process (Okpala et al., 2022; Buyl et al., 2024; Joshi et al., 2024).

A prominent form of linguistic variation is dialectal variation, characterized by macro-social factors such as geography, socioeconomic class, ethnicity, gender, and age (Haugen, 1966; Haugh and Schneider, 2012). The resulting dialects are primarily separated into two categories: standard and non-standard (Trudgill, 2004). By definition, "standard" dialects spoken by the majority receive preferential treatment from governmental and educational institutions, providing speakers of standard dialects with more socioeconomic opportunities than speakers of minority dialects (Trudgill, 1979). This creates a circular compounding effect that systemically disadvantages minority groups, leaving historically marginalized communities that speak "non-standard" language varieties vulnerable to further discrimination by modern AI systems trained on historical data (Blodgett et al., 2016; Jurgens et al., 2017; Kantharuban et al., 2023; Drożdżowicz and Peled, 2024; Nayeem and Rafiei, 2026).

Previous research has shown that components across the language modeling stack behave suboptimally on uncommon linguistic forms – such as low-resource languages and dialects – resulting in problematic biases and disparities in cost, latency, and quality that exacerbate economic divides in the accessibility of language technologies. These biases have been identified in tokenizers (Ahia et al., 2023; Petrov et al., 2023), LMs (Lin et al., 2025), and reward models (RMs; Mire et al., 2025). Systematic discrepancies in token representations of raw text across different social groups are commonly assumed to cause downstream performance issues, yet whether these biases originate from the tokenizer remains an open question.

In standard NLP pipelines, tokenization is the first step in transforming human-readable text into a machine-readable format (Grefenstette, 1999). Modern LMs use subword tokenization, which de-

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 16184–16206
October 24-29, 2026 ©2026 Association for Computational Linguistics

composes text into a sequence of subword units, or "subtokens", drawn from a tokenizer vocabulary. Crucially, while the exact algorithmic details differ, tokenizers work by optimizing a segmentation model on a training corpus, a process that does not explicitly account for uncommon linguistic forms. For example, "buildin'" might be tokenized as `['build', 'in', "'"]`, whereas "building" might be represented by the single token `['building']`. This algorithmic approach is a double-edged sword that can either improve model performance by promoting the recognition of form-based patterns (Wegmann et al., 2025) or reduce model performance by producing unwanted artifacts (Hofmann et al., 2022; Arnett and Bergen, 2024; Schmidt et al., 2024).

To advance toward fairer language technologies that minimize the undesirable effects of intra-language variation, we aim to identify the sources of dialectal bias. This work uses parallel dialect corpora in English, the highest-resource language, to trace the "dialect tax" through every stage of the modern NLP pipeline: tokenization, pre-training, post-training, and inference. By comparing English to English, our analysis removes the confounding factor of byte premiums (Arnett et al., 2024).

We begin by establishing that models recognize matched SAE and dialectal texts as semantic equivalents to a greater extent than they recognize translations or character-perturbed copies of the same text (§2). Nonetheless, our experiments document the persistence of dialectal biases across the latest generation of tokenizers and reproduce reasoning gaps across currently popular LMs (§3). Our results verify that, compared with Standard American English (SAE), dialects such as African American Vernacular English (AAVE) – which are often already stigmatized – continue to suffer from suboptimal model performance.

To investigate the origins of dialect biases, we first isolate the tokenizer's causal role using a character-level tokenization counterfactual (§4). By removing subword segmentation at inference time, we narrow the input-side dialect gap. This character-level tokenization intervention, however, leaves accuracy and output behavior largely unchanged. These results provide insight into how dialect biases persist beyond the tokenizer into the model's learned weights, necessitating further scrutiny of training dynamics. Indeed, we present empirical evidence that dialect taxes are introduced during both pre-training and post-training (§5). Across dialect pairs, we observe divergence in base-model gradients, LM hidden states, and RM preferences. Paired SAE and AAVE inputs yield gradient updates that are more dissimilar than those delivered by two entirely unrelated SAE documents, suggesting that dialectal texts consistently carry a higher prediction cost regardless of problem difficulty. RMs reward dialect-exclusive tokens in isolation yet penalize them in context, which risks propagating these biases into deployed LMs. Collectively, we show that dialectal biases originate not from tokenizers alone but accumulate across LMs and RMs. Addressing the dialect tax therefore requires new recipes for model training.

2 Same Meaning, Higher Taxes

Before tracing the source of dialect biases, we reduce concerns that observed gaps are explained by gross semantic mismatch. We verify that a text-embedding model assigns high semantic similarity to matched dialect pairs, meaning dialect-dependent behavior reflects surface-form biases.

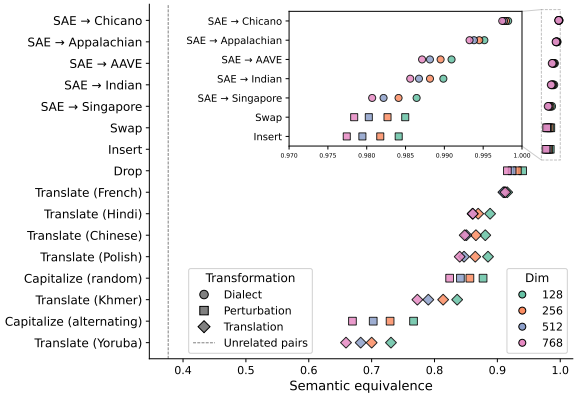


Figure 1: **Models understand semantic equivalence yet penalize surface form.** We visualize the semantic equivalence of various text transformations on MULTIVALUE. All dialect pairs achieve high similarities exceeding every perturbation and translation baseline.

Datasets. PARALLELAAVE (Groenwold et al., 2020) and MULTIVALUE (Ziems* et al., 2023) offer parallel texts between SAE and American dialects (Table 4). Each PARALLELAAVE pair contains a tweet written in AAVE and its corresponding human rewrite in SAE. Each MULTIVALUE pair uses a rule-based translation system to synthetically generate SAE texts from COQA (Reddy et al., 2018) into AAVE, Appalachian, Chicano, Indian, and Singaporean dialects.

Setup. We use EMBEDDINGGEMMA (Vera et al.,

2025), a 300M-parameter text embedding model based on the GEMMA-3 LM family, which produces numerical vector representations of four nested dimensions (768, 512, 256, 128) via Matryoshka Representation Learning (Kusupati et al., 2024).[1] We quantify semantic equivalence between two texts by calculating the cosine similarity $\text{sim}(x, y) = \langle x, y \rangle / (\|x\|_2 \|y\|_2)$ between their embeddings x and y. We compare every SAE sample in MULTIVALUE and PARALLELAAVE to three types of text transformations, which serve as a calibration baseline for changing surface form while preserving meaning to assess dialectal treatment.

Transformations. We identify a set of transformations $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_m\}$ that preserve semantic content while perturbing surface form. For each text sample x_i and transformation τ_j, we compute $\text{sim}(x_i, \tau_j(x_i))$. (1) *Dialects.* We use parallel texts. (2) *Perturbations.* We apply character-level perturbations: swaps ($\mathbb{P} = 0.05$), drops ($\mathbb{P} = 0.15$), inserts ($\mathbb{P} = 0.05$), and capitalizations (alternating or $\mathbb{P} = 0.5$). (3) *Translations.* We translate the original SAE text into six languages with high (French, Chinese), medium (Hindi, Polish), or low (Khmer, Yoruba) resources.

Dialects show semantic equivalence. We find evidence of semantic invariance in embedding models under surface-form transformations. Figure 1 compares per-sentence cosine similarities between SAE and dialect rewrites with those of SAE and perturbation or translation controls. On MULTIVALUE, all five dialect pairs achieve similarities above 0.98, significantly exceeding every perturbation and translation control baseline exhibiting similarities between 0.659 to 0.978. We confirm statistical significance (Holm-adjusted $p < 0.001$) using a one-sided paired Wilcoxon signed-rank test over 429 aligned sentences to confirm all five dialect conditions were more similar to SAE than each of the controls. On PARALLELAAVE, the SAE to AAVE similarity is 0.92, which significantly (Holm-adjusted $p < 0.001$) exceeds character deletion (0.840), capitalization (0.702 to 0.786), and translations (0.534 to 0.872), but is lower than character swapping (0.956) and insertion (0.951). All values are far above the null baseline of unrelated document pairs ($\mu = 0.41$). This floor establishes that dialect transformations, along with their original SAE texts, have scores that signal genuine similarities between each pair.

We check for the confound that the embedding model assigns higher similarities for texts that are more predictable to the model rather than for texts that preserve meaning. If the confound were true, transformations that are more surprising (higher perplexity) to a model would predict lower similarity. Instead, similarity tracks semantic content rather than LM surprisal (Table 9). Across all transformations, mean similarity correlates positively – rather than negatively – with log median per-token input-perplexity ratio (MULTIVALUE: $\rho = +0.54$, PARALLELAAVE: $\rho = +0.43$).

3 Models (Still) Have Dialectal Biases

While models have improved and hill-climbed various benchmarks, we verify that newer generations of models still exhibit token biases on information compression and downstream reasoning tasks.

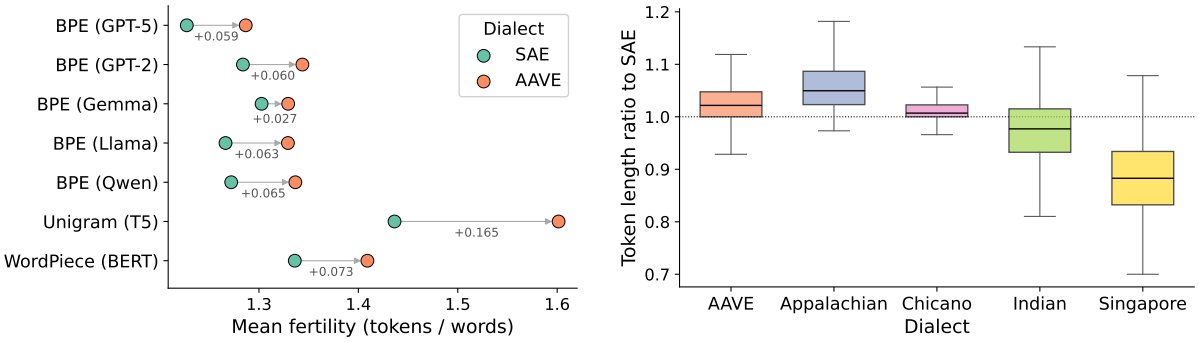
3.1 Tokenizers Have Dialectal Biases

Tokenizers. We explore the three most popular tokenization methods: byte-pair encoding (BPE) (Sennrich et al., 2015), UNIGRAM (Kudo, 2018), and WORDPIECE (Schuster and Nakajima, 2012). We evaluate tokenization metrics using the tokenizers of popular LMs (Table 5). BPE variants include GPT-5, GPT-2, GEMMA-3, LLAMA-3, and QWEN-3 models. UNIGRAM uses the T5 model. WORDPIECE uses the BERT model. We drop the LM family number where the value is obvious.

Metrics. We measure token biases through eight metrics (Table 10) and showcase the token length and fertility (number of tokens divided by the number of words) in the main section. We assess tokenizer fairness by evaluating whether these metrics achieve parity on parallel texts across all dialects, i.e. metrics should be comparable when conditioned on the same content.

Results. Tokenizers consistently produce unequal representations across English dialects. Figure 2a demonstrates that dialectal tokenization biases persist across modern tokenizer and model families. Across all tokenizers, AAVE texts have higher mean fertility (↓ better) than their SAE counterparts, with an average gap of +0.07 tokens per word. Figure 2b suggests that this tokenization bias could systematically amplify existing sociodemo-

[1] EMBEDDINGGEMMA was chosen to measure similarity, as it is explicitly optimized for tasks such as semantic similarity and is derived from one of the LM families used the analyses of later sections.



(a) Mean fertility of PARALLELAAVE

(b) Token length ratio of MULTIVALUE

Figure 2: **Modern tokenizers consistently exhibit dialectal biases.** (a) The mean fertility of the three tokenization algorithms on PARALLELAAVE shows a statistically significant difference between SAE and AAVE dialects. (b) The ratio of BPE token lengths on various dialects to that of SAE on MULTIVALUE reveals a consistent dialectal tokenization performance gap that loosely parallels current income gaps of minority groups within the US (§D.1).

graphic bias. We analyze the five non-SAE dialects in MULTIVALUE by computing the per-sample ratio of dialect token length to the paired SAE token length, which showcases a consistent tokenization bias ranking: Appalachian, AAVE, and Chicano English texts incur increased token lengths (medians above parity), while Indian and Singaporean English texts have more compact token lengths (medians below parity) than SAE. This ranking is shown in Figure 10 to be stable across all tokenizer families, indicating that the bias is structural to the dialectal text rather than an artifact of any particular algorithm. Notably, this tokenization performance gap loosely parallels the income distribution of minority groups within the US (Figure 11).

3.2 LMs Have Dialectal Biases

Setup. We modify the REDIAL (Lin et al., 2025) dataset, a reasoning benchmark containing parallel query pairs in SAE and AAVE, and reproduce their original results using a newer set of models. Changes include reformatting, instruction rewording, and correctness validation, with additional details left to Appendix A.1. We evaluate three open-source LM families (LLAMA-3: 8B, 70B; GEMMA-3: 12B, 27B; QWEN-3: 8B, 27B) and OpenAI's GPT family (GPT-5.5, GPT-5.5 MINI) to benchmark both SAE and AAVE dialects ($\sim$ 300 samples per condition) on four tasks (math, algorithm, logic, planning) using chain-of-thought ("CoT") or no CoT ("naïve") reasoning.

Results. LMs continue to show weaker reasoning capabilities on AAVE text than SAE text, as shown in Table 12. The SAE-AAVE reasoning gap exists for all model families and sizes, indicating the dialect tax remains an unsolved issue.

4 Isolating the Tokenizer from the Model

We now ask if the dialect tax is generated by subword segmentation at inference or encoded in a model's priors from training. We disentangle a *tokenizer-induced* from a *model-induced* account by forcing LMs to evaluate the same prompts under character-level tokenization that bypasses subword merges entirely. Gaps that vanish implicate the tokenizer, while gaps that persist implicate the LM.

Setup. We employ either a canonical tokenizer or a character tokenizer at inference time (§E) on nine instruction-tuned LMs across three model families (LLAMA-3: 1B, 3B, 8B; GEMMA-3: 1B, 4B, 12B; QWEN-3: 1.7B, 4B, 8B). Our method indexes off the results by Zheng et al. (2026) showing that swapping a model's canonical tokenizer retains the majority of changes in that model's behavior, and non-canonical tokenizations can outperform canonical tokenizations in some domains that require non-canonical representations of text (e.g., counting characters, arithmetic). Thus, given the known disparities between the tokenizations of dialectal text and previous findings that suboptimal tokenization leads to poor performance, character tokenization helps us isolate LM performance from tokenizer effects on SAE versus dialectal texts.

Using the REDIAL dataset, we measure: accuracy, input entropy, and output entropy. As LM behavior is necessarily entangled with its canonical tokenizer, we are primarily concerned with the relative – rather than the absolute – change in the paired dialect gap. The per-token entropy

	Llama			Gemma			Qwen		
	1B	3B	8B	1B	4B	12B	1.7B	4B	8B
Canonical (%)	99.0	99.9	99.4	99.8	99.8	100.0	99.9	100.0	100.0
Character (%)	94.5	98.3	99.6	86.4	99.5	99.5	98.5	97.3	97.5
Δ	-4.5	-1.6	$+0.2$	-13.4	-0.3	-0.5	-1.3	-2.7	-2.5

Table 1: **Dialects are linearly decodable from hidden states under both tokenizations.** We list the five-fold cross-validation accuracy of logistic regression predicting dialect from the answer-step hidden state. Character tokenization slightly reduces separability, but all models remain far above chance (50%).

gap is denoted by $\Delta_H = \mathbb{E}[H(p_\theta(\cdot|x_{<t}))]_{\text{SAE}} - \mathbb{E}[H(p_\theta(\cdot|x_{<t}))]_{\text{AAVE}}$, where $H(\cdot)$ is Shannon entropy over the predictive distribution. If the dialect tax were purely tokenizer-induced, character tokenization should close the gap on the three metrics.

Behavioral disparities remain. Dialect gaps survive forced character tokenization. These results reveal that canonical subword segmentation at inference time does not fully explain the dialectal disparity, suggesting that the dialect tax is at least partly model-induced. None of the aforementioned metrics shrink systematically under a character-level tokenizer, as pictured in Figure 3. The input entropy gap and its directionality persist under character tokenization, with AAVE positions carrying higher mean per-token input entropy than its paired SAE counterpart in 93.0% of paired samples under both tokenizations (one-sided Wilcoxon signed-rank tests, both $p < 0.001$), with 92.3% sign agreement across 19,995 pairs. The relative magnitude differences between both input entropy and output entropy of dialectal texts are mixed, suggesting that subword segmentation is not the sole contributor of the dialect tax.

Crucially, model downstream behavior is unaffected. The accuracy gap does not systematically change under character tokenization, as across nine models, five widen and four narrow, yielding statistically insignificant sign and paired t-test on gap magnitude. Our results suggest that subword segmentation inflates the gap but does not fully account for it. The output entropy gap at the answer-extraction step $H(p_\theta(\cdot|x, y_{<t^*}))$ is unchanged by the tokenizer swap across all model families. Character tokenization fails to narrow the input gap, output gap, nor accuracy disparities. This dissociation holds for individual SAE-AAVE texts, where input and output Δ_H changes are uncorrelated across $n = 13{,}197$ paired items (Pearson $r = +0.001$, $p = 0.91$). The persistence of these output gaps after bypassing subword segmentation suggests that they are attributable to the model's learned parameters rather than the tokenizer at inference time.

Representational differences remain. Our ex-

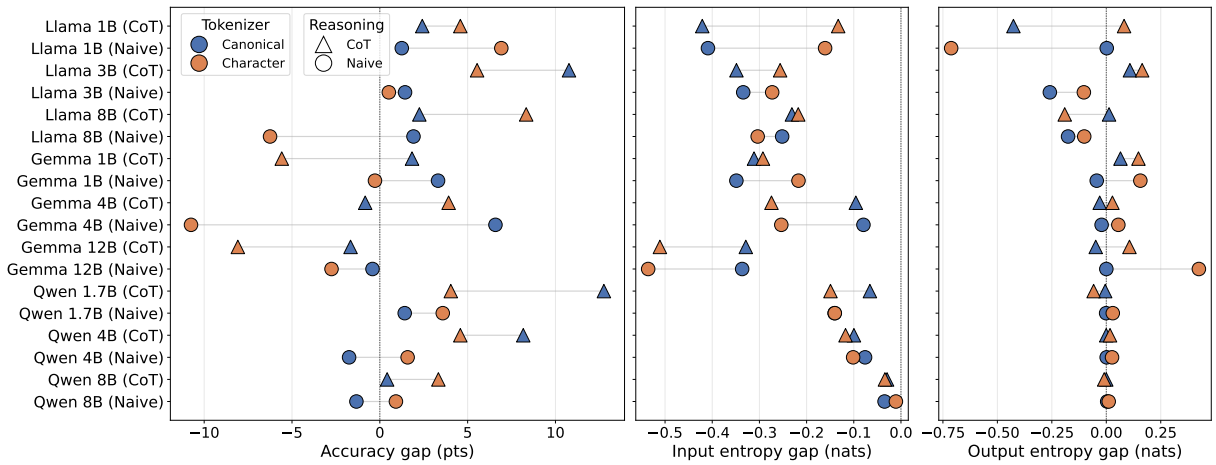


Figure 3: **Character tokenization does not systematically narrow dialect gaps.** We plot per-model SAE − AAVE effect sizes under canonical (blue) and character-level (orange) tokenization, split by reasoning strategy (△ = CoT, ◯ = naïve). Grey lines connect paired points for each model. *Left:* The accuracy gap does not systematically change under character tokenization. *Center:* The input entropy gap Δ_H changes inconsistently, confirming that character tokenization fails to equalize how the model processes dialect inputs. *Right:* The output entropy gap Δ_H persists under both tokenizations, suggesting generation-time dialect bias is independent of the tokenizer.

periments show that an LM's internal representations preserve dialect identity under character tokenization. The model-induced account predicts that the dialect should be recoverable from internal representations even after removing the tokenizer's contribution, which we test by training a logistic regression classifier to predict the dialect $d \in \{\text{AAVE}, \text{SAE}\}$ from the answer-step hidden state $\mathbf{h}_{t^*} \in \mathbb{R}^{d_{\text{model}}}$ under five-fold cross-validation. Table 1 compares the results from the classifier. Under canonical tokenization, classification accuracy is above 99% across all nine models. Under character tokenization, it remains above 86%. Even the largest drop, GEMMA 1B with -13.4 percentage points (pp), stays far above the 50% chance baseline. These results are stable across model family, scale, and reasoning strategy, supporting the claim that a model's internal representations encode dialect identity within the learned features of $\mathbf{h}_{t^*}$ independently of the subword boundaries used to produce it. Since removing canonical subword segmentation reduces but does not eliminate dialect separability, we conclude that the dialect tax is not confined to the canonical tokenizer but extends into downstream model representations.

5 Dialect Taxes Exist Despite Training

We build on the evidence that learned representations exhibit dialect taxes by exposing the same disparities in pre-training gradients (§5.1), post-training rewards (§5.2), and inference-time hidden states and output distributions (§5.3).

5.1 Pre-Training Gradients

Similar content might be expected to induce similar learning signals during training. In other words, two inputs that pose the same task and have the same target answer, differing only by a meaning-preserving surface rewrite, carry the same "knowledge" to be learned. Thus, a model that represents a task in a dialect-invariant way should produce nearly identical gradient for both; what the model learns from a document ought not to depend on the dialect in which the text is written. Under dialect-invariant learning, matched documents in different dialects should yield greater gradient alignment than unrelated documents in the same dialect. We hypothesize that matched semantic content should provide an alignment advantage after controlling for dialect. We test this hypothesis for semantically equivalent dialect pairs by measuring their per-document gradient signatures.

Setup. In accordance with previous sections, we compute the gradient projections for nine base models across three LM families (LLAMA-3: 1B, 3B, 8B; GEMMA-3: 1B, 4B, 12B; QWEN-3: 1.7B, 4B, 8B). We use the modified REDIAL dataset (§A.1) with randomized multiple-choice answers.

Metrics. Following Kazdan et al. (2026), we represent each document by the full-parameter gradient of its causal LM loss. For input x, this loss is the average next-token cross-entropy:

$$\ell(x;\theta) = \frac{1}{|x|}\sum_{u=1}^{|x|} \text{CE}(f_\theta(x)_u, x_{u+1}). \quad (1)$$

The document gradient is then:

$$g(x;\theta) = \nabla_\theta \ell(x;\theta). \quad (2)$$

Storing full gradients is impractical for billion-parameter models, so we compress each gradient via COUNTSKETCH (Spring et al., 2019), a randomized linear projection that preserves inner products in expectation. Each parameter gradient coordinate is assigned element-wise to one of $d = 8{,}192$ buckets, multiplied by a random sign, and summed with other coordinates in that bucket. This yields a sketch $\hat{g}(x;\theta) \in \mathbb{R}^d$ providing a compact approximation of the original gradient space.

We measure the cosine similarity between paired SAE-AAVE gradient projections for each sample i, $s_i^+ = \text{sim}(\hat{g}(x_i^{\text{SAE}};\theta), \hat{g}(x_i^{\text{AAVE}};\theta))$.

To calibrate these similarities, we establish a null baseline mean μ^- and standard deviation σ^- from the cosine similarity of gradient projections between unrelated SAE documents. We use this baseline to compute z-scores, $z = (\mu^+ - \mu^-)/\sigma^-$, where μ^+ is the mean paired similarity. A z-score significantly above zero indicates that dialect pairs are more similar to each other than to unrelated texts, that is, the model recognizes shared content despite dialectal variation. A z-score near zero indicates the dialect shift is as disruptive to the gradient as switching to an entirely different text.

Matched texts yield mismatched gradients. We compare two distributions: (i) the paired similarity between SAE and AAVE versions of the same text and (ii) the unpaired similarity between unrelated SAE texts. Under the null hypothesis that the model is dialect-invariant, dialect pairs that share semantic meaning should induce gradient updates

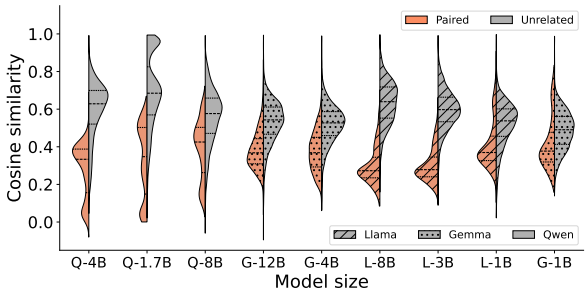


Figure 4: **Dialectal form outweighs semantic mismatch in gradient geometry.** Violin plots of the gradient cosine similarity between paired and unrelated documents on REDIAL reveal that matched SAE-AAVE pairs produce lower similarity than unrelated SAE-SAE pairs, implying that a meaning-preserving dialect shift can perturb the training signal more than changing the document content within SAE. In the plot, an LM name is indicated by its first-letter abbreviation and size.

more similar than those of unrelated documents (i.e., $\mu^+ \gg \mu^-$). Figure 4 rejects this hypothesis. Across all models and tasks, $\mu^+ < \mu^-$ with a mean $z = -2.64$, indicating that the dialect transformation disrupts gradient geometry more than substituting an entirely different document. Concretely, the gradient induced by the AAVE encoding of problem i has lower cosine similarity to that of its SAE counterpart than to that of an unrelated SAE sample j. A pre-trained base model produces more divergent parameter updates on semantically identical dialectal content than on semantically unrelated SAE content, pointing to disparities in learning costs between different dialectal texts.

Loss gaps persist regardless of correctness. The dialect loss gap is invariant to model family and scale, resulting in inflated per-token prediction cost on dialectal text. In Figure 5, all nine models assign consistently higher cross-entropy loss to AAVE inputs than to their SAE counterparts, with a mean gap of approximately 0.56 nats. Models find AAVE text systematically harder to predict than SAE text.

To rule out problem difficulty as a confound, we calculate the point-biserial correlation between s_i^+ and a binary indicator for correct answers on both dialects. Indeed, the correlation pooled across all models is negligible ($r = -0.013$, $p = 0.19$, $n = 10{,}800$), and the correlation per-model is weak ($|r| \leq 0.19$), as listed in Table 13. The conditional expectations $\mathbb{E}[s_i^+|\text{both correct}] = 0.348$ and $\mathbb{E}[s_i^+|\text{any incorrect}] = 0.354$ differ by 0.006 nats, which is much smaller than the dialect gap itself. This result falsifies the hypothesis that gradient divergence concentrates on problems the model finds difficult, meaning parameter updates are distorted uniformly regardless of correctness. Thus, improvements in model accuracy alone fail to mitigate the resulting dialect bias from pre-training.

Dialectal texts rival character noise. Naturally, one might object that any transformed text yields noisier gradients than its matched SAE text. We calibrate the AAVE z-score against five character-level perturbations[2] of SAE from Section 2, recomputing each $z = (\mu^+ - \mu^-)/\sigma^-$ where μ^+ is now the paired similarity between SAE and its perturbed version, against the shared unrelated-SAE baseline ($\mu^- = 0.57$, $\sigma^- = 0.10$). Table 2 compares the AAVE-SAE divergence to each perturbation, showing AAVE is more disruptive than character swaps or insertions and sits on the order of dropping 15% of characters. Given that AAVE preserves meaning whereas character perturbations can corrupt it, our results suggest that dialectal surface variation elicits gradient divergence comparable to that induced by substantial character-level corruption.

5.2 Post-Training Rewards

A critical component of LM alignment is the reward model (RM), which provides the optimization signal for post-training. Biases encoded by the RM propagate into the policy of the deployed LM. Here, we measure rewards on parallel SAE-AAVE content to identify systematic dialect preferences.

Setup. We evaluate ten RMs from three providers (AI2, QRM, SKYWORK) spanning 3B to 70B pa-

[2]We use swap, drop, insert, random capitalization, and alternating capitalization.

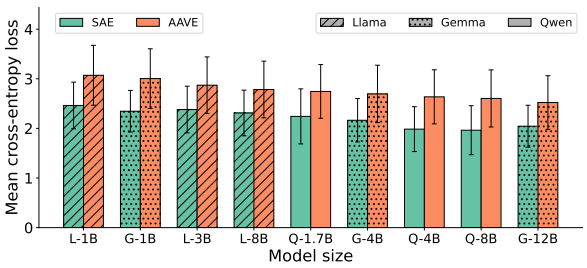


Figure 5: **Dialectal text incurs higher prediction loss.** We plot the mean cross-entropy loss by dialect. All nine models assign significantly higher loss to AAVE inputs than to their SAE counterparts (one-sided paired Wilcoxon test, $n = 1{,}200$ per model, Bonferroni-corrected $p < 0.001$), with per-model mean gaps of 0.47 to 0.66 nats (Cohen's $d_z \in [2.12, 2.55]$). An LM name is indicated by its first-letter abbreviation and size.

Paired condition with SAE	$\mu^+(\uparrow)$	z
Baseline (unrelated SAE)	0.57	0.00
Capitalize (alternating)	0.29	−3.36
Capitalize (random)	0.34	−2.88
Drop ($\mathbb{P} = 0.15$)	0.36	−2.58
Insert ($\mathbb{P} = 0.05$)	0.47	−1.40
Swap ($\mathbb{P} = 0.05$)	0.50	−1.06
Dialect (AAVE)	0.35	−2.64

Table 2: **Dialect gradient divergence sits within the range of character-level perturbations.** Across nine base models and four REDIAL tasks, we report the grand-mean paired cosine similarity (μ^+) and z-score against the unrelated-SAE baseline. Higher μ^+ indicates greater similarity to SAE gradients; lower values indicate stronger gradient divergence.

rameters across GEMMA-2, LLAMA-3, and QWEN-3 LM families (§B.4). To obtain reward scores $r(x, y)$ for an input x and response y, we calculate (1) sample-level scores and (2) token-level scores.

For sample-level scoring, we use the REDIAL dataset to construct a prompt that concatenates a reasoning problem x_i with its gold-standard answer y_i, "`{problem}{answer}`", such that each chosen-rejected preference pair presents the same question with the same answer as the assistant response in both dialects. To measure an RM preference for SAE text over AAVE text, we define $\Delta r_i = r(x_i^{\text{SAE}}, y_i) - r(x_i^{\text{AAVE}}, y_i)$.

For token-level scoring, we use REDIAL, PARALLELAAVE, and MULTIVALUE to identify subword tokens that appear exclusively in tokenized dialect texts (e.g., "wanna", "lil") and SAE texts (e.g., "Calculate", "regardless") across the seven tokenizers from Section 3. We use the method by Christian et al. (2026) to isolate the rewards of each token from any prior context. Given a fixed prompt x, "What, in one word or subword, is the greatest thing ever?", we present each token individually as the response y to receive score $r(x, y)$.

Reward models have incoherent dialect bias. If RMs scored only on semantic content, each paired SAE-AAVE reward gap should satisfy $\Delta r_i = 0$, and therefore $\mathbb{E}[\Delta r] = 0$. We test this null hypothesis with a one-sample t-test over paired reward gaps. Eight RMs reject this null at $p < 0.05$, except SKYWORK-LLAMA 3B ($p = 0.50$) and SKYWORK-GEMMA 27B ($p = 0.76$). Thus, most RMs carry some directional dialect signal, but two patterns preclude the simplest explanations of this gap. First, the bias is not a uniform AAVE penalty, as the Δr sign flips across REDIAL tasks. Pooled across RMs, Algorithm ($\overline{\Delta r} = +0.43$) and Math ($+0.04$) lean SAE while Logic ($-0.08$) and Planning ($-0.14$) lean AAVE. While all tasks are significant ($p < 0.01$) but Math ($p = 0.239$), effect sizes are uniformly small (Cohen's $|d| \leq 0.25$). That is, every RM flips dialect preference by topic. Second, the bias is not a fixed artifact of the pre-trained backbone. Within A12, the mean gap flips from $+0.41$ for the base LLAMA 8B to -0.19 after instruction tuning, reversing rather than removing the dialect tax. Taken together, these dialectal differences are neither uniform across content nor stable across training stages, rendering RM preferences too unstable to correct pre-training dialectal gaps.

Rewards are contextual, not lexical. A natural explanation for the sample-level dialect gap is that the bias lives in the surface tokens, that RMs penalize dialect-exclusive subwords. Token-level scoring reveals the opposite: dialect-exclusive tokens receive higher scores than SAE-exclusive tokens across all corpora, tokenizers, and RMs (pooled mean score gap $\bar{r}_{\text{SAE}} - \bar{r}_{\text{dialect}} = -0.55$, $p < 0.001$ with an independent t-test)[3]. The direction is consistent across each corpora (Table 16) and statistically significant with $p < 0.001$ under a pooled two-sample t-test across all five dialects after RM scaling (Table 18). Our results highlight that isolated-token and full-context preferences can be decoupled in RMs, which reinforce findings that preference-tuned LMs can sometimes exhibit covert, without overt, prejudice against AAVE (Hofmann et al., 2024). Although alignment ideally corrects undesirable pre-training biases, these RMs fail to provide consistent corrective signals for preference learning to close pre-trained dialect gaps.

5.3 Inference-Time Representations

We ask whether the dialectal biases traced above are visible during inference along two complementary axes: hidden states and output distributions.

Setup. We probe nine models across three LM families (LLAMA-3: 1B, 3B, 8B; GEMMA-3: 1B, 4B, 12B; QWEN-3: 1.7B, 4B, 8B). (1) *Hidden states.* Using both the base and instruct model variants, we extract the mean-pooled hidden state $h_\ell \in \mathbb{R}^{d_{\text{model}}}$ at every layer $\ell \in \{0, \ldots, L\}$ for each parallel pair in MULTIVALUE and PARALLELAAVE. We then compute $s_\ell = \text{sim}(h_\ell^{\text{SAE}}, h_\ell^{\text{dialect}})$ to measure the similarity of paired representations at each layer.

[3]Reward scales differ substantially across RMs, so pooled gaps are not model-balanced. We normalize this in Table 17.

(2) *Output distributions.* Using the instruct models, we compare AAVE inputs on REDIAL against the five SAE character perturbations from Section 2 by measuring their input cross-entropy, generation entropy, and answer accuracy. These perturbations calibrate AAVE against surface-form noise.

Dialects look similar to noise. All transformations produce similar hidden-layer profiles, as shown in Figure 6, that exhibit low similarity at the embedding layer due to differences in surface-form tokenizations, experience rapid convergence in similarity in intermediate layers, then diverge near the output as representations differentiate in next-token prediction. We verify that across all 504 model $\times$ dataset $\times$ transformation cells, Friedman χ^2 tests reject flat layer profiles with $p < 0.001$, and for almost every (502 of 504) condition, final-layer similarity drops significantly below its peak (one-sided Wilcoxon signed-rank test, $p < 0.001$). The magnitude of the final-layer similarities differs between transformation types. In paired comparisons across 36 model $\times$ dataset combinations, final-layer similarity for dialectal transformations is significantly higher than that for 10 of 11 comparison transformations (one-sided Wilcoxon signed-rank, Bonferroni-corrected $p < 0.01$).

In terms of entropy, dialectal transformations resemble character noise. Compared to SAE, AAVE inputs increase next-token entropy by $+0.29$ nats on average, within the range for noise-perturbed text (mean $+0.22$ nats), whereas translations produce little change (mean -0.03 nats). Generation cross-entropy remains largely unchanged under every condition ($|\Delta| < 0.03$ nats), precluding degenerate outputs for all tested text transformations.

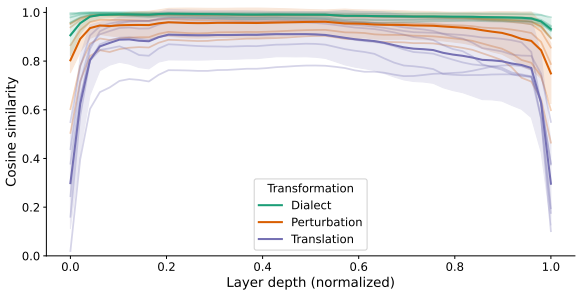


Figure 6: **Transformations result in similar hidden-state similarity curves.** Layer-wise cosine similarities between SAE hidden states and those of transformed texts are pooled across MULTIVALUE and PARALLELAAVE. All text transformations follow similar trajectories across normalized LM layers, with dialectal transformations most similar to SAE.

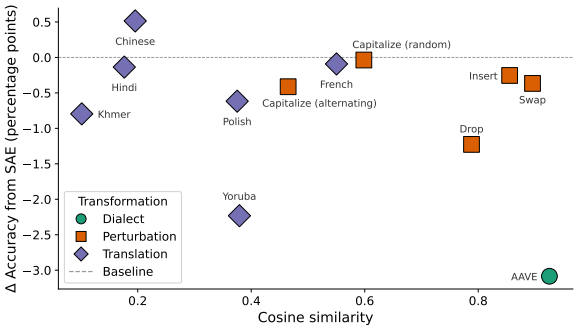


Figure 7: **Hidden-state similarity $\not\Rightarrow$ downstream accuracy.** On the REDIAL dataset, we plot each transformation's final-layer hidden-state cosine similarity to SAE against its change in answer accuracy. Among semantic-meaning-preserving transformations, SAE-to-AAVE shows the highest similarity to the original SAE representation but the lowest downstream accuracy.

Models perform worse on dialect than noise. Despite comparable internal inference-time representations, AAVE results in greater accuracy loss (3.09 pp) than noise and translation transformations (0.04 to 1.23 pp) on REDIAL (Figure 7). Ultimately, dialectal text may yield representational profiles similar to those of generic noise yet produce worse downstream outcomes, levying "the dialect tax".

6 Discussion

The key finding of this work is that the dialect tax is introduced at every stage of the language modeling pipeline. To our knowledge, this paper provides the first comprehensive study of dialect discrepancies across each component of an LM. Prior research on dialectal biases in models have examined individual language modeling components in isolation, often either surfacing the existence of a gap without identifying its source or proposing post-hoc interventions for biases already encoded in an LM. We argue that the dialect tax manifests in multiple locations and that pinpointing a singular source of dialect prejudice obscures the need to address the issue holistically. Notably, we present evidence that individual interventions mitigate specific measured disparities without resolving the underlying gap. To avoid the problem illustrated by the parable of the blind men and the elephant, proposed interventions should be evaluated for their validity across the entire model pipeline. Rather than implying that the language modeling process is prejudiced, our findings highlight the necessity of techniques that account for the systemic influence of linguistic variation throughout the model pipeline.

Limitations

Our work aims to study the sources of dialectal bias within each stage of the language modeling pipeline, but our experiments are necessarily limited by our computational resources and techniques. Below, we cover broad limitations of this study.

Datasets. Our study is empirical in nature, and as such, we needed to narrow our experimental setting to prioritize controlled parallel comparisons of standard versus non-standard English dialects. The best publicly-available resources we found differed in their construction processes. PARALLELAAVE contains human rewrites of tweets and therefore emphasizes conversational online language found on social media, REDIAL focuses on popular reasoning benchmarks written in more formal academic language, and MULTIVALUE uses a rule-based system to generate dialectal text instead of eliciting it directly from speakers. These datasets allow us to fix semantic meaning while varying dialectal surface form, which, while central to our claims, do not capture the situational range in which the dialect tax may appear. A next step is to build and evaluate community-validated parallel corpora across more dialects and settings, making sure to collect data from native speakers rather than generating text using translation rules.

Models. We host smaller open-source LMs (GEMMA-3, LLAMA-3, QWEN-3) and RMs (AI2, QRM, and SKYWORK) because their model internals (e.g., tokenizers, losses, gradients, hidden states, and reward scores) can be inspected under our computational budget. However, closed commercial systems may exhibit different behaviors, as their training pipelines may amplify or mitigate dialectal disparities in ways our methods cannot verify without internal access. We are further limited by our compute and budget constraints, which hinder our ability to repeat every experiment across multiple random seeds. Instead of measuring statistical significance across different runs, our statistical evidence relies on paired-sample variation across different inputs and models. We leave the study of dialect tax in larger models to future work.

Interventions. Our interventions diagnose mechanisms more than they prescribe a remedy. Experiments in this work are primarily observational, exposing behavioral differences to dialects at inference time but lacking controlled ablations for model training. We conduct no model training, meaning our experiments cannot identify the point at which dialectal disparities become encoded during pre-training or post-training. The intended continuation of our work aims to identify the causal mechanisms of the dialect tax through controlled experiments that explicitly test whether individual components of the language modeling pipeline preserve dialectal equivalence.

Ethics Statement

We abide by the general principles of research in the NLP community. To ensure safety and privacy, our study used open-source datasets whose data was collected with informed consent and pseudonymized participant identities.

Acknowledgments

We thank the anonymous reviewers who provided feedback for this paper. We are grateful to Valentin Hofmann and Martijn Bartelds for guidance, particularly on dataset selection, linguistic variation, and prior literature. Our gratitude extends to Joshua Kazdan and Harsha Nori for insightful discussions.

References

Judit Ács. 2019. Exploring BERT's Vocabulary.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. 2023. Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models. *Preprint*, arXiv:2305.13707.

Catherine Arnett and Benjamin K. Bergen. 2024. Why do language models perform worse for morphologically complex languages? *Preprint*, arXiv:2411.14198.

Catherine Arnett, Tyler A. Chang, and Benjamin K. Bergen. 2024. A Bit of a Problem: Measurement Disparities in Dataset Sizes Across Languages. *Preprint*, arXiv:2403.00686.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program Synthesis with Large Language Models. *Preprint*, arXiv:2108.07732.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Maarten Buyl, Alexander Rogiers, Sander Noels, Iris Dominguez-Catena, Edith Heiter, Raphael Romero, Iman Johary, Alexandru-Cristian Mara, Jefrey Lijffijt, and Tijl De Bie. 2024. Large Language Models Reflect the Ideology of their Creators. *Preprint*, arXiv:2410.18417.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating Large Language Models Trained on Code. *Preprint*, arXiv:2107.03374.

Brian Christian, Jessica A F Thompson, Elle, Vincent Adam, Hannah Rose Kirk, Christopher Summerfield, and Tsvetomira Dumbalska. 2026. Reward models inherit value biases from pretraining. In *The Fourteenth International Conference on Learning Representations*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *Preprint*, arXiv:2110.14168.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805.

Anna Drożdżowicz and Yael Peled. 2024. The complexities of linguistic discrimination. *Philosophical Psychology*, 37(6):1459–1482.

Trading Economics. 2026. Singapore GDP per capita.

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. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Gregory Grefenstette. 1999. Tokenization.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Investigating African-American Vernacular English in Transformer-Based Text Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5877–5883, Online. Association for Computational Linguistics.

Gloria Guzman. 2024. Median income of Non-Hispanic white households increased while Asian, black and Hispanic median household income did not change.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alex Wardle-Solano, Hannah Szabo, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, and 16 others. 2024. FOLIO: Natural Language Reasoning with First-Order Logic. *Preprint*, arXiv:2209.00840.

Einar Haugen. 1966. Dialect, Language, Nation. *American Anthropologist*, 68(4):922–935.

Michael Haugh and Klaus P. Schneider. 2012. Im/politeness across englishes. *Journal of Pragmatics*, 44(9):1017–1021. Im/politeness across Englishes.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633:147 – 154.

Valentin Hofmann, Hinrich Schuetze, and Janet Pierrehumbert. 2022. An Embarrassingly Simple Method to Mitigate Undesirable Properties of Pretrained Language Model Tokenizers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–393, Dublin, Ireland. Association for Computational Linguistics.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2024. Natural Language Processing for Dialects of a Language: A Survey. *Preprint*, arXiv:2401.05632.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating Dialectal Variability for Socially Equitable Language Identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57, Vancouver, Canada. Association for Computational Linguistics.

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the Dialect Gap and its Correlates Across Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7226–7245, Singapore. Association for Computational Linguistics.

Joshua Kazdan, Noam Levi, Rylan Schaeffer, Jessica Chudnovsky, Abhay Puri, Bo He, Mehmet Donmez, Sanmi Koyejo, and David Donoho. 2026. Scale dependent data duplication. *Preprint*, arXiv:2603.06603.

Taku Kudo. 2018. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. *CoRR*, abs/1804.10959.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. *CoRR*, abs/1808.06226.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2024. Matryoshka Representation Learning. *Preprint*, arXiv:2205.13147.

Kyung Ju Lee and Jeanne Batalova. 2026. Indian Immigrants in the United States.

Fangru Lin, Emanuele La Malfa, Valentin Hofmann, Elle Michelle Yang, Anthony Cohn, and Janet B. Pierrehumbert. 2024. Graph-enhanced large language models in asynchronous plan reasoning. *Preprint*, arXiv:2402.02805.

Fangru Lin, Shaoguang Mao, Emanuele La Malfa, Valentin Hofmann, Adrian de Wynter, Xun Wang, Si-Qing Chen, Michael J. Wooldridge, Janet B. Pierrehumbert, and Furu Wei. 2025. Assessing Dialect Fairness and Robustness of Large Language Models in Reasoning Tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6317–6342, Vienna, Austria. Association for Computational Linguistics.

Joel Mire, Zubin Trivadi Aysola, Daniel Chechelnitsky, Nicholas Deas, Chrysoula Zerva, and Maarten Sap. 2025. Rejected dialects: Biases against african american language in reward models. *Preprint*, arXiv:2502.12858.

Mir Tafseer Nayeem and Davood Rafiei. 2026. Which english do llms prefer? triangulating structural bias towards american english in foundation models. *Preprint*, arXiv:2604.04204.

Ebuka Okpala, Long Cheng, Nicodemus Mbwambo, and Feng Luo. 2022. AAEBERT: Debiasing BERT-based Hate Speech Detection Models via Adversarial Learning. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1606–1612.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. 2024. LogicBench: Towards Systematic Evaluation of Logical Reasoning Ability of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13679–13707, Bangkok, Thailand. Association for Computational Linguistics.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP Models really able to Solve Simple Math Word Problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language Model Tokenizers Introduce Unfairness Between Languages. *Preprint*, arXiv:2305.15425.

Alec Radford and Karthik Narasimhan. 2018. Improving Language Understanding by Generative Pre-Training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Preprint*, arXiv:1910.10683.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2018. CoQA: A Conversational Question Answering Challenge. *Preprint*, arXiv:1808.07042.

Craig W. Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. 2024. Tokenization Is More Than Compression. *Preprint*, arXiv:2402.18376.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean voice search. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural Machine Translation of Rare Words with Subword Units. *CoRR*, abs/1508.07909.

Ryan Spring, Anastasios Kyrillidis, Vijai Mohan, and Anshumali Shrivastava. 2019. Compressing gradient optimizers via count-sketches. *Preprint*, arXiv:1902.00179.

Sara Srygley, Nurfadila Khairunnisa, and Diana Elliott. 2024. The Appalachian Region: A Data Overview from the 2018-2022 American Community Survey.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

P. Trudgill. 2004. *Dialects*. Language workbooks. Routledge.

Peter Trudgill. 1979. Standard and Non-Standard Dialects of English in the United Kingdom: Problems and Policies. *International Journal of the Sociology of Language*, 1979(21):9–24.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, and 70 others. 2025. EmbeddingGemma: Powerful and Lightweight Text Representations. *Preprint*, arXiv:2509.20354.

Anna Wegmann, Dong Nguyen, and David Jurgens. 2025. Tokenization is Sensitive to Language Variation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10958–10983,

Vienna, Austria. Association for Computational Linguistics.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862, Mexico City, Mexico. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *Preprint*, arXiv:2505.09388.

Brian Siyuan Zheng, Alisa Liu, Orevaoghene Ahia, Jonathan Hayase, Yejin Choi, and Noah A. Smith. 2026. Broken tokens? your language model can secretly handle non-canonical tokenizations. *Preprint*, arXiv:2506.19004.

Caleb Ziems*, William Held*, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-VALUE: A Framework for Cross-Dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. Association for Computational Linguistics.

ZipAtlas. 2025. Immigrants from Singapore Social Profile.

Contents

1 Introduction **1**

2 Same Meaning, Higher Taxes **2**

3 Models (Still) Have Dialectal Biases **3**
3.1 Tokenizers Have Dialectal Biases . 3
3.2 LMs Have Dialectal Biases 4

4 Isolating the Tokenizer from the Model **4**

5 Dialect Taxes Exist Despite Training **6**
5.1 Pre-Training Gradients 6
5.2 Post-Training Rewards 7
5.3 Inference-Time Representations . 8

6 Discussion **9**

A Data **13**
A.1 REDIAL 13
A.1.1 Algorithm 14
A.1.2 Logic 14
A.1.3 Math 15
A.1.4 Planning 15
A.2 PARALLELAAVE and MULTI-VALUE 15
A.3 Licenses 15

B Models **15**
B.1 Tokenizers 15
B.1.1 Tokenization Algorithms . 15
B.1.2 Tokenizer Models 15
B.2 Language Models 16
B.3 Embedding Models 16
B.4 Reward Models 16
B.5 Computational Resources 16

C Semantic Equivalence **16**
C.1 Perturbations 16
C.2 Translations 17
C.3 Perplexities 17
C.4 Results on MULTIVALUE 17
C.5 Results on PARALLELAAVE . . . 17

D Benchmarks **17**
D.1 Tokenization Biases 17
D.2 Reasoning 19

E Character Tokenization **19**

F Pre-Training Analysis **19**

G Post-Training Analysis **19**
G.1 Sample-Level Rewards 19
G.2 Token-Level Rewards 19

H Inference **20**

I Language Model Usage **20**

A Data

We open source our code on GitHub at the following link:

`https://github.com/socialnlp/dialecttax`

A.1 REDIAL

Our experiments rely on the REDIAL (Lin et al., 2025) dataset, which is a reasoning benchmark containing parallel query pairs in SAE and AAVE. We reconstructed REDIAL from their original seven source datasets, which are summarized in Table 3

Category	Source	Items
Algorithm (26%)	HUMANEVAL (Chen et al., 2021)	164
	MBPP (Sanitized) (Austin et al., 2021)	149
Logic (30%)	FOLIO (Han et al., 2024; Wu et al., 2024)	162
	LOGICBENCH (Parmar et al., 2024)	200
Math (25%)	GSM8K (Cobbe et al., 2021)	150
	SVAMP (Patel et al., 2021)	150
Planning (19%)	ASYNCHOW (Lin et al., 2024)	225
Total	-	1,200

Table 3: **Source datasets used to reconstruct REDIAL.**

and detailed in the sections below. This allowed us to have greater control in modifying the natural language prompts within our experiments.

In general, we attempted to unify the format of the dataset as much as possible. For instances that conflict between REDIAL and the source datasets, we defer to the original source.

We note that there are many small inconsistencies, including punctuation, capitalization, or[4] grammar mistakes that originated from the dataset sources or were introduced by REDIAL. We corrected, to our discretion, many of these inconsistencies (described within individual source sections) and provided metadata labels to the original sources. We also remove duplicated samples.

For the experiments presented in the main section, we use our modified re-constructed REDIAL dataset formatted a multiple-choice question and answer dataset.

A.1.1 Algorithm

We remove REDIAL prompt instructions to be added programmatically during inference.

In accordance with REDIAL, we substitute all function names as `python_function` to avoid as much memorization as possible. We corrected several corrupted questions where REDIAL incorrectly replaced the original function keyword with "python_function" (e.g. we corrected "Write a function to find the nth number in the newman conway python_function." in REDIAL to "Write a function to find the nth number in the newman conway sequence.").

HUMANEVAL. The dataset includes 164 programming problems released by OpenAI. We reference the following hyperlinked code and dataset.[5] Although REDIAL modified the instructions of some samples within their data, we use the original dataset released by OpenAI.

MBPP. The MBPP dataset, a.k.a. the "Mostly Basic Python Problems" dataset, is a benchmark of crowd-sourced Python programming problems designed to be solvable by entry-level programmers. We reference the following hyperlinked dataset, specifically the hand-verified "sanitized" verison. Because MBPP did not include the context included in HUMANEVAL, we generated the context using GPT-5 MINI through OpenRouter[6] and verify the results manually. We also removed one duplicated sample within REDIAL.

A.1.2 Logic

FOLIO. The original FOLIO dataset is a manually curated logic benchmark. The counterfactual dataset was created to explore the capabilities of language models through counterfactual evaluations on various domains, such as logic. Since the perturbed counterfactual dataset contains both the original and perturbed versions, we reference the following hyperlinked code and dataset.

LOGICBENCH. The dataset offers a natural language question-answering evaluation for the logical reasoning abilities of large language models. We reference the following hyperlinked dataset. To the best of our ability, we capitalized proper nouns and the first letter of sentences to prevent confounders from non-standard grammar. The REDIAL dataset

[4] Inclusive "or", i.e. $\vee$.

[5] The REDIAL dataset used the instruction-following queries available within INSTRUCTHUMANEVAL at https://huggingface.co/datasets/codeparrot/instructhumaneval.

[6] https://openrouter.ai/

seems to have permuted the multiple choice answers, so we maintain this new ordering. We also kept the changes from REDIAL that corrected incorrect names from the original data source. See our code for more details.

A.1.3 Math

GSM8K. The dataset consists of grade-school math word problems with $8,790$ instances. We reference the following hyperlinked code and dataset. We matched each REDIAL instance to a GSM8K instance and used the original data.

SVAMP. The dataset contains $1,000$ elementary-school math word problems. We reference the following hyperlinked code and dataset. In keeping with REDIAL, we use only the SVAMP test set. To avoid unnecessary confounders in our dataset, we used GPT-5 MINI through OpenRouter to fix the grammar of the questions. We manually verify the generated results.

A.1.4 Planning

ASYNCHOW. The dataset contains benchmarks for planning reasoning tasks. We reference the following hyperlinked code and dataset. We removed 15 duplicates from the original REDIAL dataset for a total of 225 instances. Additionally, there were several inconsistencies between the original ASYNCHOW and the REDIAL datasets, e.g. missing spaces and changed characters. For any discrepancies, we defer to the original ASYNCHOW dataset. We also change the following:

- Change instructions from "These ordering constraints need to be obeyed when executing above steps" to "These ordering constraints must be followed when executing the above steps"
- Correct the capitalization of the first line description
- Correct the capitalization of label "step <#>" to "Step <#>"
- Change "Step <#>. <description>" to the more natural "Step <#>: <description>"
- Remove extraneous "\n\n" from REDIAL not present in ASYNCHOW

It is important to note that we've made the task substantially more difficult by enforcing the models to answer in seconds. This makes the task much more readily verifiable, although it is still possible to analyze the results for non-compliant completions (e.g. "3 days 7 hours") using fuzzy matching.

A.2 PARALLELAAVE and MULTIVALUE

Table 4 summarizes the dialects included in our parallel text datasets.

A.3 Licenses

REDIAL is held under the MIT License. MULTIVALUE is held under the Apache License, Version 2.0. PARALLELAAVE did not specify the license for their data, but it was published by the Association for Computational Linguistics in 2020, which defaults to the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

B Models

B.1 Tokenizers

B.1.1 Tokenization Algorithms

BPE. Based on the byte-pair encoding (BPE) compression algorithm, Sennrich et al. (2015) introduced this tokenization approach to aid in open-vocabulary machine translation tasks by encoding rare or unknown words into sequences of subword units. It is a simple technique that iteratively merges frequent pairs of bytes or byte sequences into one sequence. It is the common approach used by LM families, such as GPT (Radford and Narasimhan, 2018), GEMMA (Team et al., 2025), LLAMA (Grattafiori et al., 2024), and QWEN (Yang et al., 2025).

UNIGRAM. The algorithm is initialized with a large base vocabulary, which is iteratively trimmed to minimize the overall loss over the training data (Kudo, 2018). It is used in combination with the SENTENCEPIECE tokenizer (Kudo and Richardson, 2018) used in LMs, like T5 (Raffel et al., 2023).

WORDPIECE. Similar to BPE, WORDPIECE iteratively merges pairs of bytes or byte sequences into one sequence that maximizes the likelihood of the training data (Schuster and Nakajima, 2012). It is the tokenization algorithm used in BERT (Devlin et al., 2018).

B.1.2 Tokenizer Models

Table 5 lists the tokenizer names used in our analysis. We use the tokenizer corresponding to the largest size of a particular LM family, although this distinction does not matter.

Dataset	Source	Pairs	Dialects
PARALLELAAVE	Groenwold et al. (2020)	2019	SAE, AAVE
MULTIVALUE	Ziems* et al. (2023)	429	SAE, AAVE, Appalachian, Chicano, Indian, Singapore

Table 4: **Datasets for parallel SAE and AAVE texts.**

Tokenization	Name	Model / Encoding
BPE	GPT-5	`o200k_base` (tiktoken)
	GPT-2	`openai-community/gpt2`
	GEMMA	`google/gemma-3-27b-it`
	LLAMA	`meta-llama/Llama-3.3-70B-Instruct`
	QWEN	`Qwen/Qwen3-32B`
UNIGRAM	T5	`t5-small`
WORDPIECE	BERT	`bert-base-uncased`

Table 5: **Tokenizer names used in our analysis.**

B.2 Language Models

Table 6 lists all LMs used in our analysis, including their Hugging Face model IDs.[7]

B.3 Embedding Models

We use Google's open embedding model EMBEDDINGGEMMA-300M (Hugging Face Model ID: `google/embeddinggemma-300m`) built from GEMMA-3 to produce vector representations of text.

As suggested on the EMBEDDINGGEMMA model page, we tailor our similarity analysis query to generate embeddings that are optimized to assess text similarity by prefixing the content with the following prompt: "`task: sentence similarity | query: {content}`".

B.4 Reward Models

Table 7 lists the reward models used in our analysis.

B.5 Computational Resources

We ran our experiments on one shared compute server with 8 NVIDIA L4 GPUs and one shared compute server with 8 NVIDIA A5000 GPUs.

C Semantic Equivalence

C.1 Perturbations

Swap. We randomly replace each ASCII character with another random ASCII character. In our experiments, we swap characters with $\mathbb{P}(\text{swap}) = 0.05$.

[7]Note that there is no base model version of QWEN-3 32B.

Family	Name	Size	Type	Model ID
LLAMA	LLAMA 3.2	1B	Base	`meta-llama/Llama-3.2-1B`
	LLAMA-3.2	1B	Instruct	`meta-llama/Llama-3.2-1B-Instruct`
	LLAMA-3.2	3B	Base	`meta-llama/Llama-3.2-3B`
	LLAMA-3.2	3B	Instruct	`meta-llama/Llama-3.2-3B-Instruct`
	LLAMA-3.1	8B	Base	`meta-llama/Llama-3.1-8B`
	LLAMA-3.1	8B	Instruct	`meta-llama/Llama-3.1-8B-Instruct`
	LLAMA-3.1	70B	Base	`meta-llama/Llama-3.1-70B`
	LLAMA-3.1	70B	Instruct	`meta-llama/Llama-3.1-70B-Instruct`
GEMMA	GEMMA 3	1B	Base	`google/gemma-3-1b-pt`
	GEMMA-3	1B	Instruct	`google/gemma-3-1b-it`
	GEMMA-3	4B	Base	`google/gemma-3-4b-pt`
	GEMMA-3	4B	Instruct	`google/gemma-3-4b-it`
	GEMMA-3	12B	Base	`google/gemma-3-12b-pt`
	GEMMA-3	12B	Instruct	`google/gemma-3-12b-it`
	GEMMA-3	27B	Base	`google/gemma-3-27b-pt`
	GEMMA-3	27B	Instruct	`google/gemma-3-27b-it`
QWEN	QWEN 3	1.7B	Base	`Qwen/Qwen3-1.7B-Base`
	QWEN-3	1.7B	Instruct	`Qwen/Qwen3-1.7B`
	QWEN-3	4B	Base	`Qwen/Qwen3-4B-Base`
	QWEN-3	4B	Instruct	`Qwen/Qwen3-4B`
	QWEN-3	8B	Base	`Qwen/Qwen3-8B-Base`
	QWEN-3	8B	Instruct	`Qwen/Qwen3-8B`
	QWEN-3	32B	Instruct	`Qwen/Qwen3-32B`

Table 6: **Language models used in our analysis.**

Provider	Base Model	Size	Model ID
Skywork	LLAMA-3.2 (Instruct)	3B	`Skywork/Skywork-Reward-V2-Llama-3.2-3B`
	LLAMA-3.1 (Instruct)	8B	`Skywork/Skywork-Reward-V2-Llama-3.1-8B`
	QWEN-3 (Instruct)	4B	`Skywork/Skywork-Reward-V2-Qwen3-4B`
	QWEN-3 (Instruct)	8B	`Skywork/Skywork-Reward-V2-Qwen3-8B`
	GEMMA-2 (Instruct)	27B	`Skywork/Skywork-Reward-Gemma-2-27B`
QRM	LLAMA-3.1 (Instruct)	8B	`nicolinho/QRM-Llama3.1-8B-v2`
	GEMMA-2 (Instruct)	27B	`nicolinho/QRM-Gemma-2-27B`
Ai2	LLAMA-3.1 (Base)	8B	`allenai/Llama-3.1-8B-Base-RM-RB2`
	LLAMA-3.1 (Instruct)	8B	`allenai/Llama-3.1-8B-Instruct-RM-RB2`
	LLAMA-3.1 (Instruct)	70B	`allenai/Llama-3.1-70B-Instruct-RM-RB2`

Table 7: **Reward models used in our analysis.**

Drop. We randomly drop each character with a prespecified probability. In our experiments, we drop characters with $\mathbb{P}(\text{drop}) = 0.15$.

Insert. We randomly insert an ASCII character with a prespecified probability. In our experiments, we insert characters with $\mathbb{P}(\text{insert}) = 0.05$.

Capitalize. We have two methods of capitalizing our text: random and alternating. In the random case, we capitalize each character with a prespecified probability. In our experiments, we capitalize characters with $\mathbb{P}(\text{capitalize}) = 0.5$. In the alternating case, we alternate every other character.

C.2 Translations

We translate texts using the Google Translate API[8]. Table 8 lists the six languages we use, along with their corresponding API language codes.

Resource Level	Language	Google Translate Code
High	Chinese	`zh-CN`
	French	`fr`
Mid	Hindi	`hi`
	Polish	`pl`
Low	Khmer	`km`
	Yoruba	`yo`

Table 8: **Translation transformation details.**

C.3 Perplexities

We list input perplexities in Table 9.

C.4 Results on MULTIVALUE

Figure 8a visualizes Figure 1 from the main text and Figure 8b visualizes the corresponding semantic equivalence to Δfertility.

[8] https://cloud.google.com/translate

	MULTIVALUE		PARALLELAAVE		REDIAL	
Condition	Ratio	frac>1	Ratio	frac>1	Ratio	frac>1
Drop ($\mathbb{P} = 0.15$)	12.53	1.00	9.64	1.00	2.06	1.00
Singapore	4.92	1.00	–	–	–	–
Insert ($\mathbb{P} = 0.05$)	4.41	1.00	3.74	0.98	1.58	1.00
Indian	3.99	1.00	–	–	–	–
Drop ($\mathbb{P} = 0.05$)	3.52	1.00	3.06	0.98	1.43	1.00
AAVE	3.13	1.00	2.89	0.92	1.65	0.96
Swap ($\mathbb{P} = 0.05$)	3.08	1.00	2.33	0.94	1.36	1.00
Capitalize (random)	2.36	1.00	1.46	0.73	1.41	0.99
Appalachian	2.10	1.00	–	–	–	–
Chicano	1.46	1.00	–	–	–	–
Translate (Yoruba)	1.38	0.75	0.68	0.36	1.26	0.70
Translate (Chinese)	1.26	0.80	1.03	0.50	1.08	0.74
Capitalize (alternating)	0.73	0.37	0.38	0.14	0.92	0.40
Translate (Polish)	0.71	0.16	0.37	0.10	0.96	0.39
Translate (French)	0.64	0.08	0.34	0.05	0.96	0.34
Translate (Hindi)	0.42	0.21	0.13	0.10	0.68	0.23
Translate (Khmer)	0.24	0.15	0.05	0.11	0.38	0.17

Table 9: **Paired per-token perplexity ratio vs SAE.** We compare the perplexity ratios for character perturbations, translations, and dialects (*italicized*). Each row pairs the transformed text to its matched SAE text by `unique_id` within a model (ratio $= \exp(\text{CE}_{\text{cond}} - \text{CE}_{\text{sae}})$), aggregated as the median of per-model medians across the models listed in Table 6. The mean fraction of paired items the model finds harder than SAE is indicated under "frac>1", where a ratio above 1 means the model is more surprised by the surface form than by SAE despite preserved meaning.

C.5 Results on PARALLELAAVE

Figure 9 visualizes the same figure as Figure 8 on the PARALLELAAVE corpus.

D Benchmarks

D.1 Tokenization Biases

Metrics. Table 10 lists all token metrics used in our study. We define a "token" as an element of a tokenizer vocabulary. We define "types" as the set of unique tokens. We define a "word" (in English) as a string without any space characters. We define a "character" as any string with a length of 1.

Token Length. Figure 10 visualizes the token length ratio of various dialects to SAE on MULTI-

Metric	What does it measure?
Average tokens per word	Average number of tokens corresponding with a single real word without punctuations
Average types per word	Average number of types corresponding with a single real word without punctuations
Character length	Number of characters in the string
Fertility (Ács, 2019)	Average number of tokens corresponding with a single real word
P(in vocabulary)	Proportion of words in the tokenizer vocabulary
Token length	Number of tokens in the string
Types length	Number of types in the string
Word length	Number of words in the string

Table 10: **Tokenization metrics used to measure bias.**

VALUE for all three tokenization algorithms.

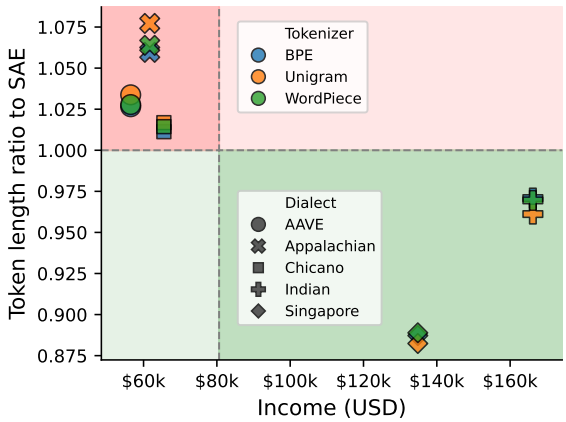


Figure 11: **Dialectal token bias and income.** We demarcate token parity with $y = 1$ and the US median household income with $x = \$80,610$. Compared to Standard American English (SAE) speakers, speakers of African American Vernacular English (AAVE), Appalachian, and Chicano dialects who earn less than the median income achieve less efficient tokenization. Meanwhile, speakers of Indian and Singaporean dialects who earn more than the median income achieve more efficient tokenization than SAE speakers.

Token bias and income. Figure 11 depicts the striking correspondence between dialectal token length ratios to SAE and the current income differences of minority groups in the US. This tokenization gap mirrors the current income gaps of minority groups within the US. Speakers of AAVE and Chicano dialects are typically black or Hispanic, respectively, who report the lowest median household incomes among all race and ethnic groups (Guzman, 2024). Median household income within the Appalachian Region is 82% of that of America overall (Srygley et al., 2024). Meanwhile, Indian Americans on average see much higher incomes than the total American population (Lee and Batalova, 2026), and Singapore is among the wealthiest countries in the world, with a GDP per capita surpassing that of America (Economics, 2026).

Table 11 details the US median household incomes (in US dollars) for different dialects. We approximate the median household income per dialect by using the median household income data of the primary demographic of the speakers of that dialect. That is, we take the median income of Black Americans to approximate the median income of AAVE speakers, and we take the median income of Hispanic Americans to approximate the median income of Chicano speakers. We use the overall median household income of the US to approximate the median income of SAE speakers.

Dialect	Income	Source
AAVE	$\$56,490$	Guzman (2024)
Appalachian	$\$61,688$	Srygley et al. (2024)
Chicano	$\$65,540$	Guzman (2024)
Indian	$\$166,200$	Lee and Batalova (2026)
SAE	$\$80,610$	Guzman (2024)
Singaporean	$\$134,818$	ZipAtlas (2025)

Table 11: **US median household income (USD).**

Token metric timeline. Figure 12 shows the chronological development of our token metrics. While we find an overall improvement in vocabulary coverage cover time, we still see evidence of the dialect tax in today's tokenizers. To quantify ranking stability, we compute Kendall's coefficient of concordance W across the seven tokenizers, treating each tokenizer as a rater of the five non-SAE dialects. Agreement is statistically significantly ($p < 0.001$) strong on every metric (fertility $W = 0.89$, $\chi^2(4) = 24.80$; vocabulary coverage $W = 0.74$; average tokens per word $W = 0.77$; average types per word $W = 0.95$; token count and word count both $W = 1.00$), indicating that the dialectal ordering is essentially fixed across tokenizers. This concordance holds despite seven years of tokenizer development. For

Model	Setting	Algorithm		Logic		Math		Planning		All		
		SAE	AAVE	SAE	AAVE	SAE	AAVE	SAE	AAVE	SAE	AAVE	Δ
GPT	CoT	93.6	90.4	80.4	76.8	94.0	92.0	96.4	92.9	91.1	88.0	**-3.09**
	Naïve	91.7	89.1	68.0	66.3	76.3	69.3	17.3	13.8	63.3	59.6	**-3.69**
GPT Mini	CoT	93.0	90.4	77.1	69.9	92.7	90.0	97.8	96.4	90.1	86.7	-3.43
	Naïve	91.4	88.2	66.9	59.9	64.0	54.0	6.7	10.7	57.2	53.2	-4.03
GEMMA 27B	CoT	83.7	80.5	66.0	66.9	92.7	89.0	24.4	24.0	66.7	65.1	**-1.62**
	Naïve	82.7	79.6	66.6	65.2	54.3	50.7	1.3	1.3	51.2	49.2	**-2.06**
GEMMA 12B	CoT	84.0	77.3	76.2	71.5	92.3	87.7	3.1	2.2	63.9	59.7	**-4.24**
	Naïve	79.9	75.7	72.4	69.6	46.3	39.7	0.0	0.4	49.6	46.4	**-3.28**
LLAMA 70B	CoT	81.2	81.2	74.3	69.9	91.3	89.3	48.0	36.9	73.7	69.3	-4.38
	Naïve	83.1	79.9	74.6	73.8	57.3	49.7	0.9	0.9	54.0	51.0	-2.92
LLAMA 8B	CoT	62.0	55.9	63.5	63.8	88.3	77.0	9.8	4.0	55.9	50.2	**-5.73**
	Naïve	58.5	55.9	66.0	58.0	41.3	31.0	0.0	0.0	41.5	36.2	**-5.23**
QWEN 27B	CoT	95.2	93.6	79.3	73.2	96.0	92.3	98.7	92.4	92.3	87.9	**-4.39**
	Naïve	55.0	54.0	73.5	67.1	70.7	68.7	2.2	0.9	50.3	47.7	-2.66
QWEN 8B	CoT	91.7	89.1	67.4	61.9	96.3	92.7	68.9	74.7	81.1	79.6	-1.49
	Naïve	69.0	65.5	47.8	43.9	34.7	33.0	0.9	0.4	38.1	35.7	-2.37

Table 12: **Benchmark results on REDIAL.** We report instruction-tuned model accuracies on REDIAL, including individual performances across the four tasks and the overall performance under the column "All". In accordance with the original REDIAL benchmark, we test for statistical significance (indicated in bold under the column "Δ") using McNemar's test for binary data and correct the p-values for multiple measurements using the Holm-Bonferroni method.

each dialect, we correlate the tokenizer's release date with its rank among the five dialects under that tokenizer (Spearman ρ, Benjamini-Hochberg FDR at $q = 0.05$). Zero of the 30 metric $\times$ dialect cells show significant rank drift, and every dialect holds a constant rank across all seven tokenizers on token count and word count. Complementary tests on the magnitude of the gap also fail to survive correction. Spearman correlations between the release date and the mean log-ratio to SAE are largest on fertility for AAVE, Appalachian, and Chicano ($\rho \approx -0.75$, $p_{\text{raw}} \in [0.02, 0.07]$), consistent with a mild improvement for those dialects, but none of the 30 comparisons survive BH correction.

D.2 Reasoning

Table 12 shows the benchmark results for REDIAL. The values track with the expected progress of LMs since the original dataset publication by Lin et al. (2025), given the modifications specified in Appendix A.1. We note that QWEN models always prefer to reason via `<think>...</think>` tokens, even on naïve prompts explicitly asking for no reasoning, so we append a model-family-specific prompt to ensure no reasoning.

E Character Tokenization

The character-tokenization experiments replace subword segmentation with a non-canonical character decomposition at inference time. This intervention keeps the model vocabulary, embedding matrix, tokenizer, and learned parameters fixed. We follow the method by Zheng et al. (2026) exactly.

F Pre-Training Analysis

Table 13 lists the point-biserial correlation between s_i^+ and correctness by base model.

G Post-Training Analysis

We list all reward models used in Table 7.

G.1 Sample-Level Rewards

Table 14 lists the per-RM dialect gap for $\overline{\Delta r}$. Table 15 denotes the sample-level rewards per task, pooled across RMs.

G.2 Token-Level Rewards

Unique words are extracted from each dataset split by whitespace and delimiter-based word segmentation. Unique tokens are extracted under each tokenizer listed in Table 5. Each word is scored as a one-word assistant response to the fixed prompt

Family	Size	r
LLAMA-3 BASE	1B	-0.119^{***}
	3B	-0.053
	8B	-0.091^{**}
GEMMA-3 BASE	1B	-0.108^{***}
	4B	-0.105^{***}
	12B	-0.066
QWEN-3 BASE	1.7B	-0.132^{***}
	4B	$+0.098^{***}$
	8B	$+0.192^{***}$
Pooled	–	-0.013

Table 13: **Point-biserial correlation between paired SAE-AAVE gradient similarity s_i^+ and a binary indicator for both-dialect correctness, by model.** Per-model sample size is $n = 1{,}200$, and pooled sample size is $n = 10{,}800$. We denote significance by *** $p < 0.001$ and ** $p < 0.01$. Pooled r is negligible, but the per-model breakdown reveals heterogeneity.

Reward model	$\overline{\Delta r}$	t
SKYWORK LLAMA 3B	-0.03	-0.67
SKYWORK QWEN 4B	$+0.32$	$+7.54^{***}$
SKYWORK LLAMA 8B	-0.22	-3.27^{**}
SKYWORK QWEN 8B	$+0.48$	$+14.42^{***}$
SKYWORK GEMMA 27B	-0.03	-0.31
QRM LLAMA 8B	$+0.01$	$+4.72^{***}$
QRM GEMMA 27B	-0.07	-3.38^{***}
AI2 LLAMA 8B Base	$+0.41$	$+12.28^{***}$
AI2 LLAMA 8B	-0.19	-10.75^{***}
AI2 LLAMA 70B	$+0.03$	$+2.06^{*}$

Table 14: **Per-RM dialect gap.** $\overline{\Delta r} = \mathbb{E}[r(x^{\text{SAE}}, y) - r(x^{\text{AAVE}}, y)]$ over REDIAL sample-level pairs ($n = 1{,}200$ per RM, i.e. 300 per task $\times$ 4 tasks), and one-sample t-statistic against $\mathbb{E}[\Delta r] = 0$. Significance is denoted by *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

"`What, in one word, is the greatest thing ever?`", and each token is scored as a one-word-or-subword assistant response to "`What, in one word or subword, is the greatest thing ever?`". For token-level dialect comparisons, for the same task and tokenizer, we identify tokens that appear in AAVE but not SAE and tokens that appear in SAE but not AAVE, then compare their isolated reward scores. Table 17 lists the per-RM per-dialect token-level dialect gap. Table 18 details the per-dialect token-level gap, before and after within-RM standardization.

H Inference

Figure 13 visualizes the hidden-state similarity dialect drops for every model on MULTIVALUE.

Task	$\overline{\Delta r}$	t
Algorithm	$+0.43$	$+13.73^{***}$
Math	$+0.04$	$+1.18$
Logic	-0.08	-2.92^{**}
Planning	-0.14	-4.92^{***}

Table 15: **Per-task dialect gap, pooled across RMs.** One-sample t-test of $\mathbb{E}[\Delta r] = 0$ within task. Significance is denoted by *** $p < 0.001$, ** $p < 0.01$.

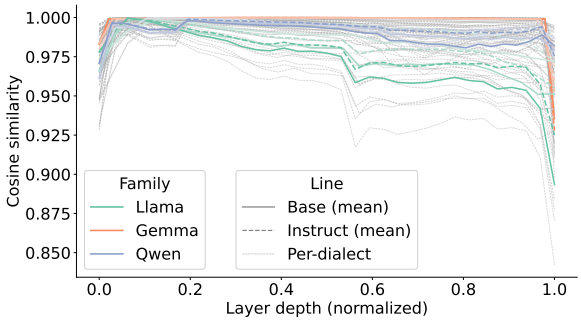


Figure 13: **Hidden-state similarity drops at the final layers for every model on MULTIVALUE.** Per-model layer-wise cosine similarity between SAE and dialect hidden states on MULTIVALUE. Grey dashed curves show each of the five dialects (AAVE, Appalachian, Chicano, Indian, Singaporean) separately; colored curves are the per-model mean across the five dialects (solid = base, dashed = instruct), with color encoding family.

Figure 14 visualizes the final-layer hidden-state similarity to SAE for each text transformation.

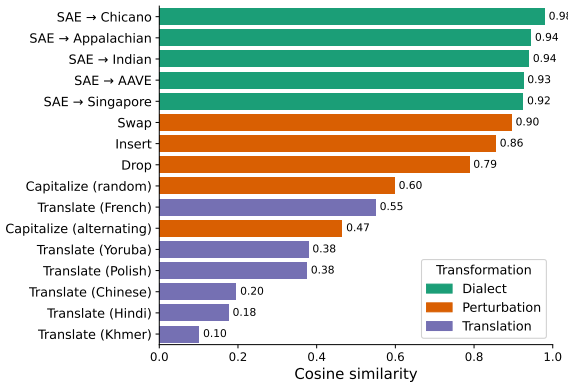


Figure 14: **Final-layer hidden-state similarity.** The final-layer hidden-state similarity to that of SAE is pooled across MULTIVALUE and PARALLELAAVE.

I Language Model Usage

We used LMs for assistance with drafting and proofreading, as well as for code completions. They served as general-purpose research tools and did not make substantive contributions to the ideation, methodology, or content of this work. The authors take full responsibility for all aspects of the work.

Corpus	Dialects	n_{SAE}	n_{dialect}	$\overline{r}_{\text{SAE}}$	$\overline{r}_{\text{dialect}}$	$\overline{r}_{\text{SAE}} - \overline{r}_{\text{dialect}}$	$\overline{d}_{\text{RM}}$	p
REDIAL	{AAVE}	121,780	163,120	-4.47	-4.10	-0.37^{***}	-0.17	9×10^{-31}
PARALLELAAVE	{AAVE}	91,380	163,420	-4.42	-3.83	-0.58^{***}	-0.29	1×10^{-67}
MULTIVALUE	{AAVE, Appal., Chic., Indian, Sing.}	141,930	132,130	-5.17	-4.74	-0.43^{***}	-0.24	9×10^{-39}
Pooled (all)		355,090	458,670	-4.74	-4.19	-0.55^{***}	-0.27	7×10^{-185}

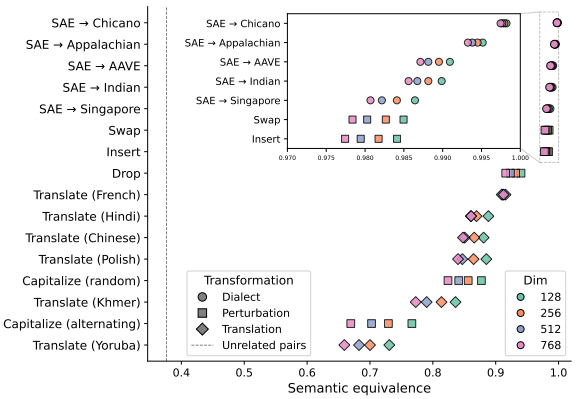
Table 16: **Per-corpus token-level dialect gap.** For each corpus, we identify subword tokens that appear exclusively in tokenized SAE vs. dialect text within each (RM, tokenizer) pairing, and score them under a fixed prompt. $\overline{r}_{\text{SAE}}$ and $\overline{r}_{\text{dial}}$ are the mean reward scores over the dialect-exclusive vocabularies. $\overline{r}_{\text{SAE}} - \overline{r}_{\text{dial}}$ is the raw gap. $\overline{d}_{\text{RM}}$ is Cohen's d computed within each RM (using that RM's pooled score standard deviation) and then averaged across the ten RMs. This normalizes for the substantial cross-RM scale differences (per-RM $\sigma \in [0.06, 3.97]$). The bottom row pools all three corpora together. Independent two-sample t-tests, with significance denoted by *** $p < 0.001$.

Reward model	AAVE	Appalachian	Chicano	Indian	Singaporean	Pooled
SKYWORK LLAMA 3B	-0.45^{***} (-0.25)	$+0.11^{*}$ ($+0.06$)	-0.71^{***} (-0.43)	-0.06 (-0.03)	-0.88^{***} (-0.49)	-0.51^{***} (-0.28)
SKYWORK QWEN 4B	-0.69^{***} (-0.31)	-0.32^{***} (-0.15)	-0.71^{***} (-0.37)	-0.13^{*} (-0.06)	-1.10^{***} (-0.47)	-0.78^{***} (-0.34)
SKYWORK LLAMA 8B	-1.26^{***} (-0.38)	-0.09 (-0.03)	-1.49^{***} (-0.49)	-0.28^{**} (-0.08)	-1.34^{***} (-0.38)	-1.26^{***} (-0.37)
SKYWORK QWEN 8B	-0.80^{***} (-0.40)	-0.16^{**} (-0.09)	-0.67^{***} (-0.35)	$+0.00$ ($+0.00$)	-0.61^{***} (-0.29)	-0.73^{***} (-0.36)
SKYWORK GEMMA 27B	-1.04^{***} (-0.27)	-1.74^{***} (-0.45)	-1.97^{***} (-0.54)	$\mathbf{+2.15}^{***}$ ($\mathbf{+0.57}$)	-2.08^{***} (-0.52)	-1.17^{***} (-0.30)
QRM LLAMA 8B	$+0.00^{***}$ ($+0.04$)	-0.00 (-0.01)	-0.02^{***} (-0.37)	-0.00 (-0.05)	-0.01^{***} (-0.19)	-0.00^{*} (-0.02)
QRM GEMMA 27B	-0.15^{***} (-0.30)	-0.18^{***} (-0.37)	-0.23^{***} (-0.46)	-0.12^{***} (-0.24)	-0.17^{***} (-0.34)	-0.16^{***} (-0.33)
A12 LLAMA 8B Base	-0.42^{***} (-0.26)	-0.39^{***} (-0.25)	-1.05^{***} (-0.74)	-0.38^{***} (-0.22)	-1.19^{***} (-0.71)	-0.55^{***} (-0.34)
A12 LLAMA 8B	-0.09^{***} (-0.12)	-0.07^{***} (-0.10)	-0.28^{***} (-0.47)	-0.07^{***} (-0.09)	-0.30^{***} (-0.43)	-0.13^{***} (-0.18)
A12 LLAMA 70B	-0.16^{***} (-0.15)	-0.02 (-0.03)	-0.13^{**} (-0.17)	-0.05^{*} (-0.05)	-0.22^{***} (-0.22)	-0.18^{***} (-0.17)
Mean across RMs	-0.50 (-0.24)	-0.29 (-0.14)	-0.73 (-0.44)	$+0.11$ (-0.03)	-0.79 (-0.40)	-0.55 (-0.27)

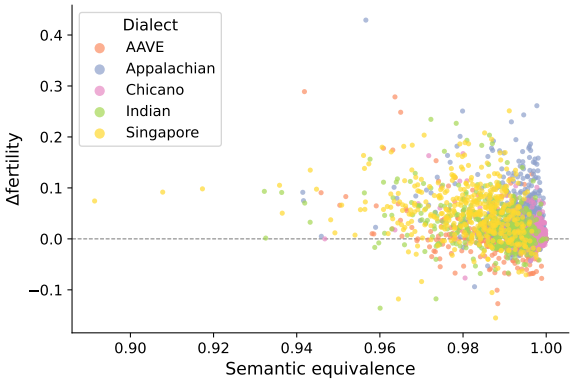
Table 17: **Per-RM per-dialect token-level dialect gap.** Each cell shows the raw reward gap $\overline{r}_{\text{SAE}} - \overline{r}_{\text{dial}}$ with Cohen's d in parentheses. Negative values indicate the RM scores dialect-exclusive tokens higher than SAE-exclusive ones. AAVE-exclusive tokens come from REDIAL, PARALLELAAVE, and MULTIVALUE, and other dialect-exclusive tokens come from MULTIVALUE. Cohen's d uses each RM's pooled score standard deviation as denominator, so raw gaps are not directly comparable across RMs while d values are. The bolded cell is the only one flipping the dialect-favoring pattern. The "Pooled" column aggregates a given RM across all five dialects, and the "Mean across RMs" row is the unweighted average across the ten RMs. We run independent two-sample t-tests and denote the significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

			Raw scores		Within-RM standardized	
Dialect	n_{SAE}	n_{dial}	$\overline{r}_{\text{SAE}} - \overline{r}_{\text{dial}}$	p	d	p
AAVE	239,000	351,140	-0.50	6×10^{-115}	-0.24	$< 10^{-300}$
Appalachian	24,980	31,540	-0.29	8×10^{-5}	-0.13	5×10^{-64}
Chicano	8,960	6,640	-0.73	3×10^{-7}	-0.39	2×10^{-165}
Indian	27,200	32,710	$+0.11$	0.14	-0.03	4×10^{-4}
Singaporean	54,950	36,640	-0.79	6×10^{-40}	-0.40	$< 10^{-300}$
All pooled	355,090	458,670	-0.55	7×10^{-185}	-0.27	$< 10^{-300}$

Table 18: **Per-dialect token-level gap, before and after within-RM standardization.** *Raw scores:* pooled two-sample t-test on raw reward scores. Note that per-RM output scales differ, so the test is heavily influenced by the high-σ Skywork models. *Within-RM standardized:* each RM's scores are z-scored using its own mean and standard deviation, then pooled across RMs before the t-test. The d column reports the gap in standard deviations, which is mathematically equivalent to the unweighted mean Cohen's d across the ten RMs (within-RM effect size).

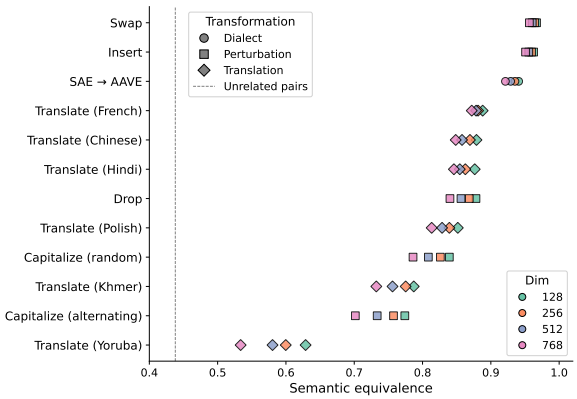


(a) Semantic equivalence of transformations

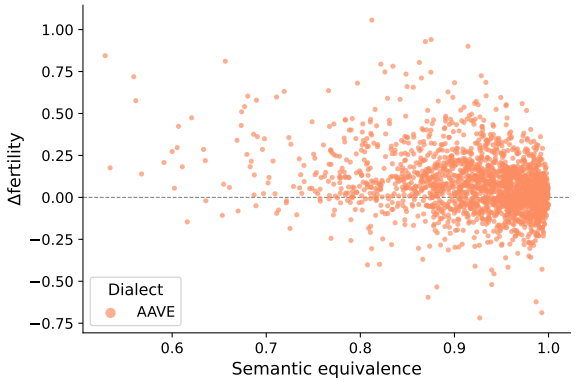


(b) Semantic equivalence to Δfertility

Figure 8: **Models understand semantic equivalence yet penalize surface form.** We find evidence of semantic invariance under surface-form transformations, as shown on the MULTIVALUE corpus. (a) We visualize the semantic equivalence of various text transformations. All dialect pairs achieve high cosine similarities above 0.97, which exceed every perturbation and translation baseline. (b) We plot semantic equivalence (cosine similarity at $d = 768$) against the tokenization tax ($\Delta\text{fertility} = \mathbb{E}[\text{fertility}_{\text{dialect}}] - \mathbb{E}[\text{fertility}_{\text{SAE}}]$) for each dialect-SAE pair. Samples in the upper-right exhibit high semantic equivalence but higher tokenization cost, where meaning is preserved while a tax is imposed.



(a) Semantic equivalence of transformations



(b) Semantic equivalence to Δfertility

Figure 9: **Models understand semantic equivalence yet penalize surface form.** We find evidence of semantic invariance under surface-form transformations, as shown on the PARALLELAAVE corpus. (a) We visualize the semantic equivalence of various text transformations. The AAVE pairs achieve high cosine similarities, which exceed most perturbation and translation baselines. (b) We plot semantic equivalence (cosine similarity at $d = 768$) against the tokenization tax ($\Delta\text{fertility} = \mathbb{E}[\text{fertility}_{\text{AAVE}}] - \mathbb{E}[\text{fertility}_{\text{SAE}}]$) for each AAVE-SAE pair. Samples in the upper-right exhibit high semantic equivalence but higher tokenization cost, where meaning is preserved while a tax is imposed.

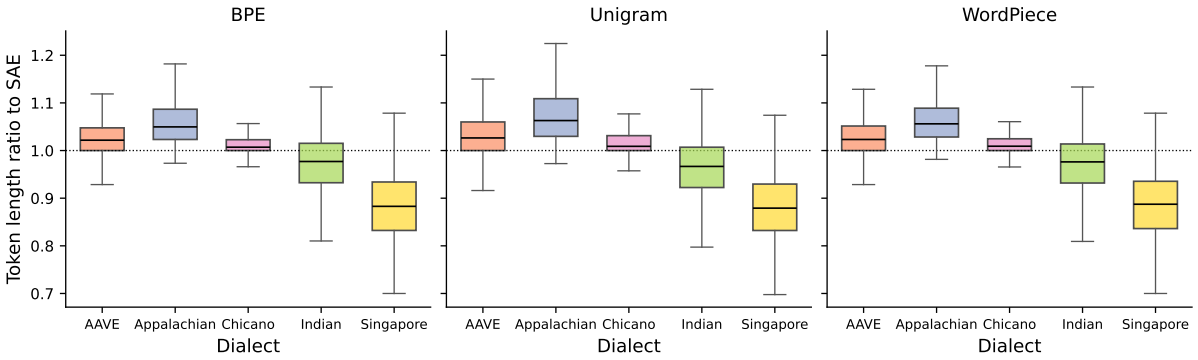


Figure 10: **Ratio of dialect to SAE token lengths on MULTIVALUE.** Dotted lines indicate token-length parity with the paired SAE text. We see similar tokenization bias ranking on every measured metric for the six dialects in the dataset (Appalachian > AAVE > Chicano > SAE > Indian > Singapore). While some dialects (AAVE, Appalachian, Chicano) have increased token lengths compared to SAE, other dialects (Indian, Singapore) have decreased token lengths compared to SAE. We find that the same dialectal token bias remains present across all three tokenization strategies.

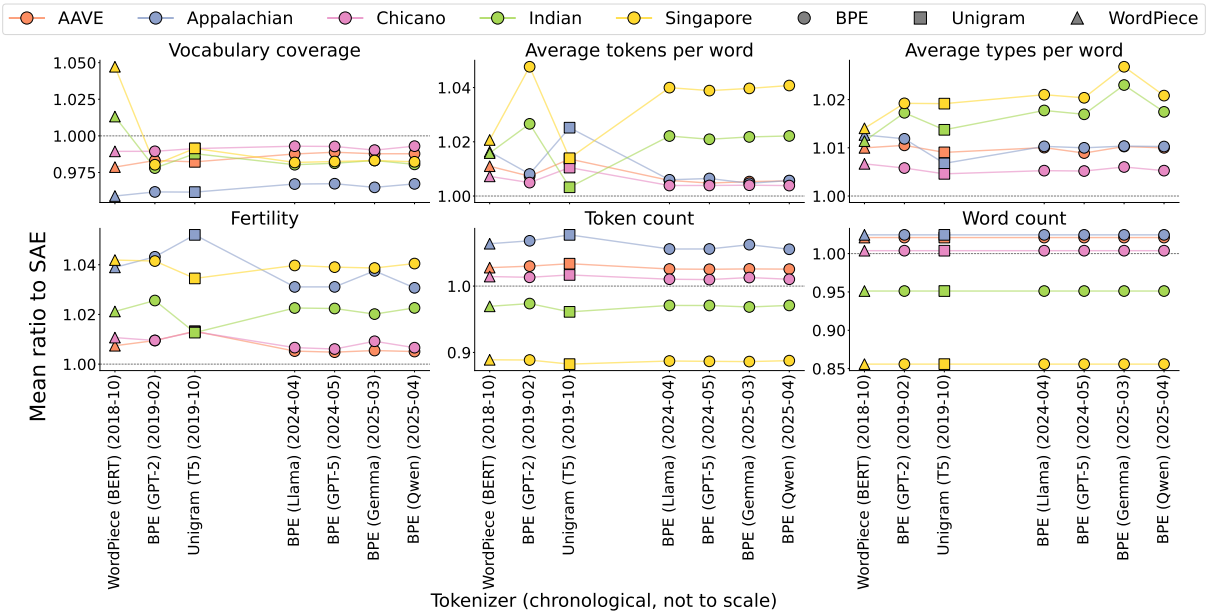


Figure 12: **Tokenizer metric timeline.** We plot a subset of the token metrics in Table 10 over time.