EnDive: A Cross-Dialect Benchmark for Fairness and Performance in Large Language Models

Anonymous ACL submission

Abstract

The diversity of human language, shaped by social, cultural, and regional influences, presents significant challenges for natural language processing (NLP) systems. Existing benchmarks often overlook intra-language variations, leaving speakers of non-standard dialects underserved. To address this gap, we introduce **EnDive** (**En**glish **Div**ersity), a benchmark that evaluates seven state-of-the-art (SOTA) large language models (LLMs) across tasks in language understanding, algorithmic reasoning, mathematics, and logic. Our framework translates Standard American English datasets into five underrepresented dialects using few-shot prompting with verified examples from native speakers, and compare these translations against rule-based methods via fluency assessments, preference tests, and semantic similarity metrics. Human evaluations confirm high translation quality, with average scores of at least 6.02/7 for faithfulness, fluency, and formality. By filtering out near-identical translations, we create a challenging dataset that reveals significant performance disparities—**models consistently underperform on dialectal inputs compared to Standard American English (SAE).** **EnDive** thus advances dialect-aware NLP by uncovering model biases and promoting more equitable language technologies.

1 Introduction

Language diversity, shaped by social and cultural factors, presents significant challenges for NLP systems. While English serves as a global lingua franca, its dialects exhibit substantial variation that often goes unaddressed in language technologies (Chambers and Trudgill, 1998). This oversight perpetuates discrimination against dialect speakers in critical domains like education and employment (Purnell et al., 1999; Hofmann et al., 2024a), exacerbated by LLMs' predominant focus on SAE (Blodgett et al., 2016).

Recent studies reveal systemic biases in LLM processing of non-standard dialects (Fleisig et al., 2024; Resende et al., 2024)—from toxic speech misclassification of African American Vernacular English tweets (Sap et al., 2019) to parsing errors in Chicano and Jamaican English (Fought, 2003; Patrick, 1999). Similar issues plague Indian and Singaporean English due to morphological divergences (Kachru, 1983; Gupta, 1994), highlighting an urgent need for inclusive NLP systems (Ziems et al., 2022).

Existing benchmarks like GLUE (Wang et al., 2019) and SuperGLUE (Wang et al., 2020) fail to capture dialect variation, while specialized datasets (SVAMP, MBPP, FOLIO) (Patel et al., 2021; Austin et al., 2021; Han et al., 2024) remain SAE-centric. Recent large-scale efforts such as DIALECTBENCH (Faisal et al., 2024) and AraDiCE (Mousi et al., 2024) broaden coverage to hundreds of varieties or to Arabic dialects, but they still leave cross-dialect reasoning largely unexplored. While frameworks like Multi-VALUE (Ziems et al., 2023) address dialect representation through rule-based lexical substitutions, their synthetic approach fails to capture authentic syntactic patterns. This limitation is particularly acute in reasoning tasks, where surface-level translations preserve logical meaning but lose dialect-specific pragmatic markers essential for fair evaluation.

To address these gaps, we introduce **EnDive** (**En**glish **Diversity**), a benchmark that evaluates five LLMs across 12 natural language understanding (NLU) tasks translated into five underrepresented dialects selected for their linguistic distinctiveness and sociocultural significance:

- **African American Vernacular English (AAVE)**: 33M speakers with distinct syntax/phonology (Lippi-Green, 1997)
- **Indian English (IndE)**: 250M speakers blending local/colonial influences (Kachru, 1983)

- **Jamaican English (JamE)**: Diaspora language with mesolectal variation (Patrick, 1999)
- **Chicano English (ChcE)**: Spanish-influenced variety in US Hispanic communities (Fought, 2003)
- **Colloquial Singaporean English (CollSgE)**: Multicultural creole with Asian substrates (Platt and Weber, 1980)

Our methodology combines linguistic authenticity with strategic filtering to create robust dialect evaluations. Using verified text samples in the target dialects from eWAVE (Kortmann et al., 2020) for few-shot prompting, we translate SAE datasets into target dialects while preserving sociolinguistic nuance. To eliminate superficial transformations, we apply BLEU-based filtering (Papineni et al., 2002), removing translations with scores ≥ 0.7 against their SAE sources—retaining only substantive linguistic variations that challenge LLMs' dialect understanding. We compare our translations against Multi-VALUE's rule-based translations (Ziems et al., 2023) through fluency assessments, semantic similarity metrics, and LLM preference tests. Additionally, we have native speakers assess our translations to ensure linguistic authenticity and original content meaning are preserved across all five dialects.

Our Contributions:

(1) Public Benchmark: Curated challenging dialectal variants across 12 reasoning and natural language understanding tasks, validated via multiple metrics and human evaluation.

(2) Cross-LLM Evaluation: We evaluated seven SOTA models using chain-of-thought (CoT) and ZS prompting to assess performance disparities between SAE and dialectal inputs.

2 Related Work

Dialectal Diversity. Addressing dialectal diversity in NLP remains a significant challenge due to inherent linguistic variations shaped by social and cultural contexts. Early research identified systemic biases in language models against non-standard dialects such as AAVE, highlighting issues like the misclassification of AAVE tweets as toxic and difficulties in syntactic parsing (Sap et al., 2019; Jørgensen et al., 2015). Recent studies extend these findings to modern LLMs, revealing persistent dialect prejudice in evaluations related to employability, criminality, and medical diagnoses (Hofmann et al., 2024b; Fleisig et al., 2024; Blodgett and O'Connor, 2017).

Sociolinguistic Impact and Real-World Discrimination. Beyond technical benchmarks, sociolinguistic studies have linked LLM biases to real-world discrimination—such as housing denials for AAVE speakers (Hofmann et al., 2024b; Purnell et al., 1999) and biased criminal justice assessments (Fleisig et al., 2024). Multilingual initiatives like LLM for Everyone (Cahyawijaya, 2024) advocate for continuously fine-tuning models to better serve underrepresented languages. Our approach reflects this tuning perspective by using human-guided few-shot prompting with authentic linguistic examples (Kortmann et al., 2020; Platt and Weber, 1980) to generate dialect-specific translations that effectively "tune" the input data, ensuring that the unique features of underrepresented dialects are accurately captured. This alignment helps mitigate model biases and promotes more equitable language technologies.

Benchmarking Approaches and Hybrid Methodologies. Dialect robustness is primarily evaluated using two approaches. The first relies on rule-based lexical substitutions—exemplified by VALUE and Multi-VALUE (Ziems et al., 2022, 2023)—which are scalable but often miss nuanced, context-dependent features (e.g., AAVE's habitual "be" (Green, 2002; Lippi-Green, 1997) or Chicano English's Spanish-influenced prosody (Fought, 2003; Santa Ana, 1993). The second employs human-annotated or community-driven translations (e.g., ReDial; AraDiCE (Lin et al., 2025; Mousi et al., 2024); CultureBank (Shi et al., 2024). Recent hybrid methodologies combine automated generation with native-speaker validation, as in CulturePark (Li et al., 2024) for cross-cultural dialogue and AraDiCE for Arabic. Complementary evaluation sets such as CulturalBench (Chiu et al., 2024) measure everyday cultural knowledge rather than dialect syntax, highlighting a parallel but related gap. Meanwhile, AAVENUE (Gupta et al., 2024) provides human-validated AAVE benchmarks. These hybrid approaches offer a more robust framework for comprehensive dialect and culture fairness evaluations.

Remaining Gaps and Our Contribution. Although prior work has deepened our understanding of dialect biases in NLP, significant gaps remain in developing comprehensive, multi-dialect benchmarks that integrate authentic linguistic features. ENDIVE addresses these gaps by providing a robust benchmark that combines both automated and

human-validated translation methods, thereby fostering more equitable language technology development.

3 Dataset

3.1 Dataset Overview

ENDIVE is a benchmark designed to evaluate the reasoning capabilities of LLMs across five underrepresented dialects. The benchmark is curated from 12 established datasets, spanning four core reasoning categories: **Language Understanding**, **Algorithmic Understanding**, **Math**, and **Logic**. Tasks were translated from SAE into the target dialects using few-shot prompting informed by eWAVE examples. For comparison, we generate parallel translations using Multi-VALUE's rule-based framework.

3.2 Data Sourcing

The dataset comprises tasks selected from diverse, established benchmarks. For every benchmark we **randomly sampled** a subset of instances to keep the overall benchmark tractable while preserving topic diversity. Below we list each source dataset, its focus, and the number of examples drawn.

Language Understanding **BoolQ** (Wang et al., 2020) is a yes/no question-answering task derived from Wikipedia passages that measures factual consistency; we randomly sampled 1,000 instances. **MultiRC** (Wang et al., 2020) requires multi-sentence reasoning where each question may have multiple correct answers; we randomly sampled 1,000 examples. **WSC** (Wang et al., 2020) evaluates commonsense coreference resolution by asking a model to link pronouns to their correct referents; we randomly sampled 659 examples. **SST-2** (Wang et al., 2019) is a binary sentiment-classification benchmark based on movie reviews; we randomly sampled 1,000 instances. **COPA** (Wang et al., 2020) presents a premise and two alternatives, asking the model to choose the more plausible cause or effect; we randomly sampled 500 examples.

Algorithmic Understanding **HumanEval** (Chen et al., 2021) consists of Python programming problems accompanied by unit tests; we randomly sampled 164 examples. **MBPP** (Austin et al., 2021) contains beginner-friendly Python tasks for program synthesis and correctness checking; we randomly sampled 374 examples.

Math **GSM8K** (Cobbe et al., 2021) comprises grade-school math word problems that require multi-step numeric reasoning; we randomly sampled 1,000 examples. **SVAMP** (Patel et al., 2021) offers systematically perturbed arithmetic problems designed to test robustness in mathematical reasoning; we randomly sampled 700 examples.

Logic **LogicBench** (Parmar et al., 2024) evaluates deductive reasoning through both Yes/No and four-choice formats; we randomly sampled 980 total items—500 Yes/No and 480 multiple-choice. **FOLIO** (Han et al., 2024) frames first-order-logic challenges in natural language and asks models to judge truth or contradiction; we randomly sampled 1,000 examples for this task.

3.3 Few-Shot Prompting for Dialect Translation

To translate tasks from SAE into each of the five underrepresented dialects, we employed a few-shot prompting strategy (Brown et al., 2020) informed by examples from eWAVE (Kortmann et al., 2020), a linguistically validated resource that documents and analyzes structural variations across global English dialects. We utilized three utlized exemplar translations from eWAVE per dialect. Utilizing GPT-4o (OpenAI, 2024a), the language model was then prompted to rewrite the input text in the desired dialect based on these exemplars. This approach ensures that translations maintain linguistic authenticity and accurately reflect the sociocultural nuances inherent to each dialect. Detailed examples of these prompts can be found in Appendix H.

3.4 Comparison with Rule-Based Translations from Multi-VALUE

To evaluate the effectiveness of our human-guided few-shot prompting method, we compare our dialectal translations against those generated by Multi-VALUE (Ziems et al., 2023). Multi-VALUE is a rule-based framework that applies predefined linguistic rules to transform SAE into target dialects in a systematic manner. This comparison allows us to assess how well our approach captures authentic dialectal variations relative to a purely rule-based method.

The percentage of successful translations for each dataset and dialect is detailed in Appendix A, where we observe that Multi-VALUE often failed to return valid outputs due to SpaCy errors produced by its tool. This inconsistency underscores

Dataset	AAVE	IndE	JamE	CollSgE
BoolQ	**0.8326** / 0.6202	**0.8080** / 0.7757	**0.7785** / 0.5456	**0.7145** / 0.6062
COPA	**0.7076** / 0.6833	**0.7659** / 0.5633	0.6391 / **0.3633**	**0.7074** / 0.5947
MultiRC	**0.8239** / 0.5626	**0.7982** / 0.7728	**0.8151** / 0.4793	**0.7325** / 0.5160
SST-2	**0.7985** / 0.5777	**0.7634** / 0.7285	**0.7786** / 0.4650	0.7005 / **0.5941**
WSC	**0.7488** / 0.6503	0.6540 / **0.3594**	**0.7341** / 0.4013	**0.6298** / 0.6069
HumanEval	N/A / N/A	**0.8993** / 0.7854	**0.8265** / 0.6238	N/A / N/A
MBPP	**0.8188** / 0.7617	**0.8853** / 0.7297	**0.7370** / 0.6289	**0.7088** / 0.6181
GSM8K	**0.8079** / 0.7055	**0.8006** / 0.7543	**0.7784** / 0.5263	**0.6698** / 0.6553
SVAMP	**0.8038** / 0.7498	**0.8418** / 0.7632	**0.7896** / 0.5346	**0.6980** / 0.6661
Folio	**0.7737** / 0.6492	**0.8474** / 0.7607	**0.7787** / 0.5805	**0.6920** / 0.6475
Logic Bench MCQ	**0.7847** / 0.4953	**0.8841** / 0.7421	**0.7808** / 0.4541	**0.6751** / 0.4447
Logic Bench YN	**0.4742** / 0.2183	**0.8139** / 0.7401	**0.7788** / 0.4386	**0.6732** / 0.4331
Average	**0.7613** / 0.6067	**0.8135** / 0.7063	**0.7679** / 0.5034	**0.6911** / 0.5802

Table 1: *ROUGE Diversity Scores across Dialects and Datasets (***ENDIVE** */ Multi-VALUE).* Bold indicates the higher score.

Dataset	AAVE	IndE	JamE	CollSgE
BoolQ	**-1.84** / -2.05	**-1.08** / -2.10	-3.92 / **-2.21**	-2.52 / **-2.45**
COPA	**-2.26** / -3.08	**-1.65** / -2.97	-5.65 / **-2.94**	-3.53 / **-3.38**
MultiRC	-2.29 / **-2.00**	**-1.14** / -2.24	-4.41 / **-2.03**	-2.86 / **-2.29**
SST-2	-3.21 / **-2.96**	**-2.39** / -3.73	-5.18 / **-3.30**	-4.09 / **-3.49**
WSC	**-2.14** / -2.78	**-1.23** / -2.87	-4.98 / **-2.49**	**-2.88** / -3.39
HumanEval	N/A / N/A	**-2.80** / -3.13	-3.53 / **-2.46**	N/A / N/A
MBPP	**-1.65** / -2.51	**-1.25** / -3.31	-4.17 / **-3.09**	**-2.83** / -3.20
GSM8K	**-1.82** / -2.06	**-1.12** / -2.27	-4.06 / **-2.31**	**-2.35** / -2.87
SVAMP	**-1.74** / -2.28	**-1.16** / -2.33	-4.02 / **-2.45**	**-2.34** / -3.11
Folio	**-2.16** / -2.48	**-1.21** / -2.57	-3.54 / **-2.47**	**-2.89** / -2.96
Logic Bench MCQ	-2.53 / **-2.24**	**-1.09** / -2.42	-4.50 / **-2.27**	-3.08 / **-2.92**
Logic Bench YN	-2.55 / **-2.46**	**-1.21** / -2.48	-4.53 / **-2.31**	-3.09 / **-2.99**
Average	**-2.20** / -2.45	**-1.44** / -2.70	-4.37 / **-2.53**	**-2.95** / -3.00

Table 2: *BARTScores across Dialects and Datasets (***ENDIVE** */ Multi-VALUE).* Scores closer to 0 indicate better performance. Bold indicates the better (less-negative) score.

the need for more robust and context-aware translation methods, such as our few-shot prompting approach with GPT-4o, which consistently generates fluent and faithful dialectal rewrites.

3.5 BLEU Score Filtering for Challenging Translations

To create a more challenging benchmark, we applied BLEU score (Papineni et al., 2002) filtering to exclude translations with sentence-level BLEU above 0.70, as these were nearly identical to the original SAE text. This retained examples with greater linguistic diversity and surface variation, emphasizing authentic dialectal shifts. The 0.70 threshold was chosen empirically based on score distributions (Appendix B) to balance semantic alignment with structural divergence. This was especially important for dialects like AAVE and JamE, where subtle edits can mask deeper grammatical changes.

4 Analysis

4.1 ROUGE Diversity Evaluation

ROUGE Diversity (Lin, 2004), computed as the average of ROUGE-1, ROUGE-2, and ROUGE-L, captures lexical richness while maintaining semantic fidelity. As shown in Table 1, **ENDIVE** outperforms Multi-VALUE across all four dialects, with the largest margin in **JamE**—highlighting its strength in capturing diverse and expressive phrasing in more structurally distinct varieties. These results suggest that EnDive introduces greater surface-level variation while maintaining semantic

Each metric (BARTScore, ROUGE, and BLEU) is computed by comparing the dialectal translation to its original SAE version. This allows us to assess semantic similarity, surface overlap, and fluency preservation relative to the original input.

Dataset	AAVE	IndE	JamE	ChcE	CollSgE
BoolQ	6.51	6.41	6.11	6.05	5.88
COPA	6.83	6.39	6.55	6.27	5.41
MultiRC	6.83	6.03	6.01	6.01	5.96
SST-2	6.64	5.84	5.85	5.93	5.58
WSC	6.36	5.97	5.50	6.15	5.60
HumanEval	6.12	6.44	6.45	6.35	6.26
MBPP	6.01	6.71	5.62	6.10	5.28
GSM8K	6.37	6.29	6.15	6.38	6.10
SVAMP	6.14	6.18	5.69	6.21	5.71
FOLIO	6.74	5.82	6.06	6.26	5.93
Logic Bench MCQ	6.35	5.75	6.21	6.28	5.76
Logic Bench YN	6.38	5.60	6.24	6.22	5.79
Average	**6.44**	**6.12**	**6.04**	**6.18**	**5.77**

Table 3: *Fluency Scores for* **ENDIVE** *Translations Across Datasets and Dialects (1–7).* Higher scores indicate better fluency as evaluated by GPT-4o.

Dataset	IndE	AAVE	CollSgE	JamE
BoolQ	**100.00** / 0.00	**100.00** / 0.00	**100.00** / 0.00	**100.00** / 0.00
COPA	**95.22** / 4.78	**95.80** / 4.20	**95.69** / 4.31	**98.07** / 1.93
MultiRC	**100.00** / 0.00	**100.00** / 0.00	**100.00** / 0.00	**100.00** / 0.00
SST-2	**95.15** / 4.85	**97.99** / 2.01	**97.86** / 2.14	**98.05** / 1.95
WSC	**100.00** / 0.00	**99.25** / 0.75	**100.00** / 0.00	**99.28** / 0.72
HumanEval	**97.34** / 2.66	N/A / N/A	N/A / N/A	**100.00** / 0.00
MBPP	**100.00** / 0.00	**99.53** / 0.47	**99.70** / 0.30	**100.00** / 0.00
GSM8K	**99.75** / 0.25	**99.71** / 0.29	**99.78** / 0.22	**99.63** / 0.37
SVAMP	**100.00** / 0.00	**98.66** / 1.34	**99.02** / 0.98	**98.01** / 1.99
FOLIO	**99.32** / 0.68	**98.19** / 1.81	**99.67** / 0.33	**99.31** / 0.69
Logic Bench MCQ	**99.12** / 0.88	**100.00** / 0.00	**99.78** / 0.22	**100.00** / 0.00
Logic Bench YN	**100.00** / 0.00	**100.00** / 0.00	**99.58** / 0.42	**99.76** / 0.24
Average	**98.82** / 1.18	**99.01** / 0.99	**99.19** / 0.81	**99.34** / 0.66

Table 4: *Preference Scores for* ***Claude 3.5 Sonnet*** *Across Datasets and Dialects (***ENDIVE** */ Multi-VALUE).* N/A indicates no valid preferences. Bold indicates the better score.

integrity, highlighting its ability to generate fluent and dialectally rich text.

4.2 BARTScore Evaluation

BARTScore (Yuan et al., 2021) is a learned metric of generation quality where values closer to 0 (i.e. less negative) indicate better outputs. As Table 2 shows, ENDIVE outperforms Multi-VALUE in three of four dialects—AAVE, IndE (the largest improvement), and CollSgE—while JamE remains the primary challenge. This highlights that **ENDIVE** generates high-quality, dialect-sensitive outputs across the board, though JamE shows that Multi-VALUE can still be competitive in certain cases.

4.3 Lexical Diversity Evaluation

Lexical diversity captures how varied the vocabulary is in a model's output, reflecting its ability to adapt to dialect-specific expressions. As shown in Appendix D, **ENDIVE** consistently outperforms Multi-VALUE across all four dialects, with the largest margin in IndE (0.8087 vs. 0.7284). These results highlight ENDIVE's strength in producing more varied and expressive generations while preserving meaning, reinforcing its fluency and adaptability across dialects.

4.4 Fluency Evaluation

Fluency measures how natural and grammatically correct a translation sounds in its target dialect. Since most automatic fluency metrics are tuned for SAE, we follow prior work (Kocmi and Federmann, 2023) and use **GPT-4o** (OpenAI, 2024a) to evaluate fluency via CoT prompting (Appendix J). As shown in Table 3, **ENDIVE** performs fluently across all dialects, with especially high scores in AAVE and ChcE, suggesting that its generations

Rubric Item	Multi-VALUE	ENDIVE
Accurate & consistent AAVE grammar	All young teenage girls at attends musics festival frequently big fans of pop bands and singers.	All young teenage girls who **be hittin'** up music festivals all the time **is** real into pop bands and singers.
AAVE contractions (ain't, gon')	If a movie popular, some person enjoy watching it.	If a movie **poppin'**, some folks like **watchin'** it. All things that some folks enjoy **gon'** get attention.
AAVE conversational vocabulary	All red fruits that which is growing in Ben's yard are containing some Vitamin C.	All **da** red fruits **growin'** in Ben's yard **got** some Vitamin C.
AAVE syntactic structure (zero copula)	All social mediums applications containing chat features are softwares.	All social media apps with chat features, **they software**.

Table 5: AAVE examples. **BrickRed** = core AAVE grammatical features (habitual *be*, zero-copula "they software," tense/aspect markers like *gon'*, vernacular lexemes *da*, *got*). **MidnightBlue** = phonological spellings or contractions typical of AAVE (*poppin', watchin', growin'*).

Rubric Item	Multi-VALUE	ENDIVE
JamE grammar (articles "di," plural "dem," relativiser "weh")	All citizens of Lawton Park are using the a zip a code 98199.	All **di people dem weh** live **inna** Lawton Park use **di** zip code 98199.
JamE contractions / plural marker	All fruits that is growing in Ben's a yard and are containing some A Vitamin A C are healthy.	All **di fruit dem weh** grow **inna** Ben yard an' have some Vitamin C **a** good **fi yuh**.
JamE conversational vocabulary	If Nancy is not toddler, then Nancy is seafarer.	If Nancy **nuh likkle pickney**, **den** Nancy **a** seafarer.
JamE negative particle "nah"	If someone young, then they are not elderly.	If somebody young, **den dem nah** elderly.
JamE omission of articles / auxiliaries	Functional brainstems are necessary for breath control.	Functional brainstems **necessary fi** control **yuh** breath.

Table 6: JamE examples. **BrickRed** = core Patois morpho-syntactic markers (article *di*, plural *dem*, relativiser *weh*, preverbal marker *a*, negative *nah*, focus particle *den*). **MidnightBlue** = phonological/lexical elements and locatives typical of JamE (*inna*, *fi yuh*, pronoun *yuh*).

are both readable and dialect-consistent.

4.5 Preference Evaluation

To assess overall translation quality, we conducted pairwise preference tests comparing **ENDIVE** to **Multi-VALUE**, using CoT prompting to evaluate fluency, accuracy, readability, and cultural fit (see Appendix K). As shown in Table 4, **ENDIVE** was consistently preferred across all dialects, achieving average win rates above 98% in AAVE, IndE, JamE, and CollSgE. Even in the lowest-margin case—CollSgE COPA—**ENDIVE** maintained a clear lead (73.92%). Additional results in Appendix D show similar preferences under **GPT-4o** and **Gemini 1.5**, further confirming that our translations better align with dialect-specific expectations.

4.6 Qualitative Analysis

ENDIVE consistently produces more authentic and contextually appropriate dialectal translations than rule-based Multi-VALUE, which often substitutes isolated words without capturing broader syntactic or cultural patterns.

To make these differences visually transparent, we color-code key linguistic markers: **BrickRed** marks core morpho-syntactic features (e.g., habitual *be* in AAVE, plural *dem* in JamE, passive *kena* in CollSgE), while **MidnightBlue** flags contractions, phonological spellings, and discourse particles (e.g., *poppin'*, *inna*, *ah*). This helps reviewers see at a glance what Multi-VALUE misses and how **ENDIVE** addresses it.

For **AAVE**, **ENDIVE** uses habitual "be" (**be hittin' up**), colloquial forms (**gon'**), and particles (**da**), as seen in Table 5.

In **JamE**, Table 6 highlights plural markers (**di people dem**), relativiser "weh," and negation (**nah**), structures Multi-VALUE fails to replicate.

For **ChcE**, **ENDIVE** captures relaxed syntax and progressive forms (**be writin'**), avoiding ungrammatical outputs like *goed*.

Model	AAVE		ChcE		CollSgE		IndE		JamE	
	CoT	*SAE CoT*	*CoT*	*SAE CoT*	*CoT*	*SAE CoT*	*CoT*	*SAE CoT*	*CoT*	*SAE CoT*
Gemini 2.5 Pro	88.89	**92.06**	88.70	**92.31**	89.02	**92.14**	89.72	**92.24**	89.19	**92.18**
o1	89.13	**93.15**	88.54	**93.39**	89.14	**93.50**	90.34	**94.07**	89.40	**93.14**
Claude 3.5 Sonnet	79.78	**83.10**	81.15	**88.78**	81.15	**88.83**	79.61	**88.82**	80.18	**88.79**
GPT-4o	82.20	**87.36**	80.37	**87.35**	82.43	**87.31**	83.30	**87.34**	82.53	**87.44**
DeepSeek-v3	82.06	**87.36**	81.55	**87.27**	81.65	**87.37**	82.90	**87.44**	81.40	**87.38**
GPT-4o-mini	74.53	**78.27**	75.01	**77.70**	80.59	**86.61**	74.26	**86.63**	80.56	**86.60**
LLaMa-3-8B Instruct	82.69	**87.49**	78.08	**82.94**	78.41	**83.00**	81.52	**86.12**	79.14	**83.20**
Average	82.75	**86.97**	81.91	**87.11**	83.20	**88.39**	83.09	**88.95**	83.20	**88.39**

Table 7: *Average CoT accuracy (%) across 12 tasks for seven models and five dialects, with the bottom row showing the column-wise means.* Each bolded value is the higher score between CoT and SAE CoT for that dialect. Full model evaluation tables are in Appendix E.

In **IndE**, BrickRed marks local constructions (**are being**, **only**) while MidnightBlue tags culturally grounded terms (**rupees**, **paise**), missing in Multi-VALUE.

CollSgE features sentence-final particles (**lah**, **ah**, **siah**) and omitted auxiliaries (**kena** passive), shown in Table 25.

See Appendix 4.6 for extended, color-annotated examples of ChcE, IndE, and CollSgE generations.

4.7 Human Validators

Dialect	Faithfulness	Fluency	Formality	Info Retention
AAVE	6.28	6.28	6.28	6.63
ChcE	6.40	6.33	6.26	6.71
IndE	**6.45**	**6.62**	**6.59**	**6.91**
JamE	6.37	6.28	6.33	6.66
CollSgE	6.19	6.11	6.02	6.52
Average	6.34	6.32	6.30	6.69

Table 8: Native speaker evaluation scores (1-7 scale).

To validate translation quality, we conducted human evaluations with native speakers of each dialect assessing 120 randomly sampled translations. Evaluators rated outputs on four key dimensions using 7-point Likert scales (1=worst, 7=best): *Faithfulness*, ensuring the translated text conveys identical semantic content as the original; *Fluency*, guaranteeing grammatical correctness and natural flow in the target dialect; *Formality*, maintaining appropriate register and sociolinguistic conventions; and *Information Retention*, verifying no factual information was lost during the dialect conversion process. Our annotators were all native speakers who had grown up in communities where the target dialect is predominantly spoken, each holding at least a college degree to ensure language fluency and evaluative rigor. Annotators were recruited through academic networks to ensure trusted participation. These evaluations confirmed that our translations successfully maintain linguistic authenticity while preserving original content meaning, stylistic conventions, and factual integrity across all dialects. Detailed scores are shown in Table 8.

SAE ZS and SAE CoT are included for each dialect because datasets are filtered via BLEU scores, meaning each dialect has a different dataset composition. This ensures a fair comparison of model performance across dialectal variations.

5 Results and Discussion

We evaluate seven SOTA LLMs across five dialects and twelve tasks: **Gemini 2.5 Pro** (DeepMind, 2025), **o1** (OpenAI, 2024b), **Claude 3.5 Sonnet** (Anthropic, 2024), **GPT-4o** and **GPT-4o-mini** (OpenAI, 2024a), **DeepSeek-v3** (DeepSeek-AI, 2024), and **LLaMa-3-8B Instruct** (META, 2024). As shown in Table 7, we report each model's average accuracy under CoT prompting on both SAE and dialectal inputs.

Performance Gaps Between SAE and Dialectal Inputs. All seven models demonstrate consistent performance drops when evaluated on dialectal inputs compared to SAE prompts. The average gap ranges from approximately 2.69% to 12.37%, with smaller models like **GPT-4o-mini** showing the steepest decline—most notably a 12.37% drop on IndE (74.26% CoT vs. 86.63% SAE CoT). In contrast, top-tier models like **o1** exhibit greater resilience, with gaps typically under 5% across all dialects. This consistent disparity across five dialects underscores the presence of systematic bias, even under reasoning-augmented prompting strategies.

Reasoning-Only Models Lead, but Gaps Re-

Dialectal Question	SAE Equivalent	Gold	Predicted	Model(s)
The committee done approved the bill, right?	Did the committee approve the bill?	Yes	No	Claude 3.5 Sonnet, o1
Ain't no one gone to that party, huh?	Did anyone go to that party?	No	Yes	Claude 3.5 Sonnet, LLaMa-3-8B Instruct
She been had that promotion, right?	Has she had that promotion?	Yes	No	GPT-4o-mini
He don't work on Mondays, right?	Does he work on Mondays?	No	Yes	DeepSeek-v3, Claude 3.5 Sonnet

Table 9: Representative semantic misalignments in yes/no QA: dialectal questions, their SAE equivalents, the correct label, the model's (incorrect) prediction, and which models exhibit the error.

main. o1 and **Gemini 2.5 Pro**, both reasoning-native models, deliver the strongest performance across all dialects and prompting conditions. **o1** consistently scores above 88.5% on dialectal inputs and surpasses 93% under SAE CoT prompting. **Gemini 2.5 Pro** shows similar strength, with dialectal scores ranging from 88.70% to 89.72% and SAE CoT scores consistently above 92%. Despite their robust performance, both models still exhibit performance gaps of 3–5 points between dialectal and SAE inputs. This indicates that architectural advances and reasoning capabilities alone are insufficient to fully close the dialect gap.

Mid-Tier Models Show Mixed Robustness. GPT-4o, **DeepSeek-v3**, and **Claude 3.5 Sonnet** form a middle tier in performance, generally scoring between 79%–83% on dialectal inputs. **Claude 3.5 Sonnet** has scores ranging from 79.61% on IndE to 81.15% on CollSgE, and performance gaps exceeding 9 points in multiple dialects. These patterns suggest that while mid-sized models benefit from CoT prompting, they remain vulnerable to dialectal shifts, especially in settings involving complex linguistic variation.

Smaller Models Are More Sensitive to Dialect Shift. GPT-4o-mini and **LLaMa-3-8B Instruct**, the two smallest models in the benchmark, consistently yield lower accuracies and exhibit wider gaps between dialectal and SAE inputs. **GPT-4o-mini** records 74.53% on AAVE compared to 78.27% with SAE CoT, while **LLaMa-3-8B Instruct** shows similarly notable drops—particularly on ChcE and CollSgE. These findings underscore the disproportionate impact of dialectal variation on smaller or less instruction-tuned models.

Dialectal Disparities Persist Across Model Tiers. Despite strong aggregate performance from top models, consistent accuracy gaps across dialects reveal a clear and persistent issue: models struggle to generalize beyond Standard American English. Mid-sized and smaller systems are especially vulnerable, but even the best-performing models exhibit measurable degradation across dialects. This points to a central conclusion—current language models, regardless of scale, **exhibit dialectal bias**. Full per-task results are provided in Appendix E.

5.1 Dialect-Induced Errors

A common failure mode in yes/no QA is semantic misalignment—where models misinterpret polarity in dialectal inputs. Constructions like double negatives ("ain't no one"), habitual aspect ("don't be"), or markers like "been had" often cause models to flip the correct answer.

Table 9 shows cases where dialectal phrasing led to incorrect judgments, despite clear SAE equivalents. These errors appeared frequently in BoolQ and affected models across tiers, including Claude 3.5 and GPT-4o-mini.

Such failures show that dialectal bias extends beyond syntax—it disrupts semantic understanding. Even large multilingual models struggle with the pragmatic structure of non-standard English. More examples are in Appendix G.

6 Conclusion

This paper introduces **ENDIVE**, a benchmark for evaluating LLMs on dialectal robustness across 12 diverse NLP tasks and five underrepresented English dialects. Our results show that LLMs consistently underperform on non-standard dialects compared to SAE, highlighting significant unfairness and limitations in current language technologies. Moving forward, we aim to expand **ENDIVE** to additional dialects and refine translation methodologies to further bridge the gap in dialect-aware NLP. By establishing this benchmark, we encourage future research into fairer, more robust intra-language technologies that serve all linguistic communities.

7 Limitations

ENDIVE evaluates LLM performance across 12 reasoning tasks spanning four categories, using queries adapted from well-established benchmarks. While these tasks capture key reasoning challenges, they do not cover all aspects of dialectal variation, and additional task types such as Figurative Language Understanding, Commonsense Reasoning, and Conversational Reasoning may reveal further biases.

Furthermore, we tested five widely used LLMs. However, given the rapid pace of development in the field, it is infeasible to evaluate every emerging model. We hope **ENDIVE** will serve as a resource for future studies examining fairness and robustness across a broader range of LLMs as they emerge.

We also acknowledge that while GPT-4o was used to generate dialectal translations, it underperforms on dialectal reasoning tasks. Nonetheless, we distinguish between generation and comprehension: GPT-4o, when prompted with verified few-shot examples, produces fluent and faithful outputs, as validated by native speakers (Section 4.7). Still, LLM-generated text may not fully capture deeper linguistic and cultural nuances. Comparing these translations to naturally occurring dialect corpora would further strengthen our evaluation of realism.

We faced limitations with BLEU Score filtering as well. For ChcE, the number of remaining translations was extremely low because Multi-VALUE struggled to generate diverse translations and many were further filtered out due to BLEU score thresholds. As a result, there were too few data points to evaluate ChcE translations against Multi-VALUE. A similar issue arose with HumanEval for AAVE and CollSgE, where limited translations prevented reliable evaluation of metrics for these dialects.

Finally, while our results highlight significant performance disparities in dialectal inputs, this study does not deeply investigate the underlying causes of these discrepancies or propose direct mitigation strategies. Understanding these biases and developing equitable NLP solutions remain important areas for future research. Despite these limitations, we believe **ENDIVE** provides a valuable framework for advancing dialect-aware NLP evaluation.

8 Ethics Statement

We recognize the ethical considerations involved in evaluating LLM biases through the **ENDIVE** benchmark and have taken steps to ensure ethical data collection, recruiting and evaluation.

For data collection, **ENDIVE** utilizes few-shot prompting with examples from eWAVE to generate dialectal translations. While this provides systematic and scalable translations, we recognize it does not fully capture the depth of dialectal variation. We do not claim to capture the full depth of any dialect, and we encourage further work that incorporates human-validated translations for a more nuanced representation. Additionally, we were mindful to avoid reinforcing stereotypes or misrepresentations in dialect translations.

For our human validators, we recruited fluent native speakers from diverse dialect communities to ensure our translations accurately reflect cultural and linguistic nuances. Validators were fairly compensated for their contributions and encouraged to take breaks to avoid fatigue, ensuring quality and well-being throughout the process. We also do not collect personal information from validators.

Moreover, our evaluation combines LLM-based assessments with human validation to mitigate model bias. However, we acknowledge that LLMs may still reflect inherent biases, and our benchmark does not yet address the root causes of these disparities.

Despite these limitations, **ENDIVE** aims to advance equitable NLP development and encourages ongoing research to enhance dialect representation in language models.

References

Anthropic. 2024. Claude-3.5-sonnet technical report. Website.

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.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial

disparity in natural language processing: A case study of social media african-american english. *Preprint*, arXiv:1707.00061.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Samuel Cahyawijaya. 2024. Llm for everyone: Representing the underrepresented in large language models. *Preprint*, arXiv:2409.13897.

J.K. Chambers and Peter Trudgill. 1998. *Dialectology*, 2nd edition. Cambridge University Press, Cambridge, UK.

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, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024. Culturalbench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms. *Preprint*, arXiv:2410.02677.

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.

Google DeepMind. 2025. Gemini model and thinking updates: March 2025. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/. Accessed: 2025-05-16.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. Dialectbench: A nlp benchmark for dialects, varieties, and closely-related languages. *Preprint*, arXiv:2403.11009.

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. Linguistic bias in chatgpt: Language models reinforce dialect discrimination. *Preprint*, arXiv:2406.08818.

Carmen Fought. 2003. *Chicano English in Context*. Palgrave Macmillan, New York, USA.

Lisa J. Green. 2002. *African American English: A Linguistic Introduction*. Cambridge University Press, Cambridge, UK.

Abhay Gupta, Philip Meng, Ece Yurtseven, Sean O'Brien, and Kevin Zhu. 2024. Aavenue: Detecting llm biases on nlu tasks in aave via a novel benchmark. *Preprint*, arXiv:2408.14845.

Anthea Fraser Gupta. 1994. *The Step-Tongue: Children's English in Singapore*. Multilingual Matters, Clevedon, UK.

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, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander R. Fabbri, Wojciech Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. 2024. Folio: Natural language reasoning with first-order logic. *Preprint*, arXiv:2209.00840.

Valentin Hofmann, Pratyusha R. Kalluri, Dan Jurafsky, et al. 2024a. Ai generates covertly racist decisions about people based on their dialect. *Nature*, 633:147–154.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024b. Dialect prejudice predicts ai decisions about people's character, employability, and criminality. *Preprint*, arXiv:2403.00742.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 9–18, Beijing, China. Association for Computational Linguistics.

Braj B. Kachru. 1983. *The Indianization of English: The English Language in India*. Oxford University Press, Delhi, India.

Tom Kocmi and Christian Federmann. 2023. Gemba-mqm: Detecting translation quality error spans with gpt-4. *Preprint*, arXiv:2310.13988.

Bernd Kortmann, Kerstin Lunkenheimer, and Katharina Ehret, editors. 2020. *eWAVE*. eWAVE.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024. Culturepark: Boosting cross-cultural understanding in large language models. *Preprint*, arXiv:2405.15145.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Fangru Lin, Shaoguang Mao, Emanuele La Malfa, Valentin Hofmann, Adrian de Wynter, Xun Wang, Si-Qing Chen, Michael Wooldridge, Janet B. Pierrehumbert, and Furu Wei. 2025. One language, many gaps: Evaluating dialect fairness and robustness of large language models in reasoning tasks. *Preprint*, arXiv:2410.11005.

Rosina Lippi-Green. 1997. *English with an Accent: Language, Ideology, and Discrimination in the United States*. Routledge, London & New York.

META. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2024. Aradice: Benchmarks for dialectal and cultural capabilities in llms. *Preprint*, arXiv:2409.11404.

OpenAI. 2024a. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024b. Introducing openai o1. `https://openai.com/o1/`. Accessed: 2025-05-16.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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. *Preprint*, arXiv:2404.15522.

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.

Peter L. Patrick. 1999. *Urban Jamaican Creole: Variation in the Mesolect*. John Benjamins Publishing, Amsterdam, Netherlands.

John Platt and Heidi Weber. 1980. *English in Singapore and Malaysia: Status, Features, Functions*. Oxford University Press, Singapore.

Thomas Purnell, William Idsardi, and John Baugh. 1999. Perceptual and phonetic experiments on american english dialect identification. *Journal of Language and Social Psychology*, 18(1):10–30.

Guilherme H. Resende, Luiz F. Nery, Fabrício Benevenuto, Savvas Zannettou, and Flavio Figueiredo. 2024. A comprehensive view of the biases of toxicity and sentiment analysis methods towards utterances with african american english expressions. *Preprint*, arXiv:2401.12720.

Otto Santa Ana. 1993. Chicano english and the nature of the chicano language setting. *Hispanic Journal of Behavioral Sciences*, 15(1):3–35.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Chunhua yu, Raya Horesh, Rogério Abreu de Paula, and Diyi Yang. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. *Preprint*, arXiv:2404.15238.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. Superglue: A stickier benchmark for general-purpose language understanding systems. *Preprint*, arXiv:1905.00537.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. *Preprint*, arXiv:1804.07461.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Preprint*, arXiv:2106.11520.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. VALUE: Understanding dialect disparity in NLU. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3701–3720, Dublin, Ireland. Association for Computational Linguistics.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-value: A framework for cross-dialectal english nlp. *Preprint*, arXiv:2212.08011.

A Multi-VALUE Completed Translations

Dataset	AAVE (%)	ChcE (%)	CollSgE (%)	IndE (%)	JamE (%)
BoolQ	100.0	35.5	41.7	41.9	42.0
COPA	100.0	45.8	100.0	100.0	97.0
Folio	100.0	76.9	90.0	89.6	89.7
GSM8K	100.0	85.7	95.0	95.0	95.0
HumanEVAL	100.0	11.6	11.6	11.6	11.6
Logic Bench MCQ	100.0	100.0	100.0	100.0	100.0
Logic Bench Yes/No	100.0	100.0	100.0	100.0	100.0
MBPP	100.0	39.8	99.7	99.7	99.2
MultiRC	100.0	43.3	47.8	48.9	49.1
SST-2	100.0	96.3	96.3	96.2	96.3
SVAMP	100.0	74.7	93.2	93.2	93.0
WSC	100.0	73.9	92.7	92.8	92.9

Table 10: Percentage of Translations Successfully Completed by Multi-VALUE Across Dialects and Datasets

B BLEU Score Filtering Statistics

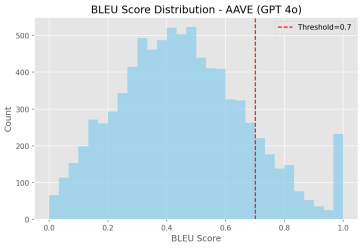
Dataset	AAVE (%)	ChcE (%)	CollSgE (%)	IndE (%)	JamE (%)
BoolQ	7.59	0.50	2.00	59.96	0.40
COPA	15.40	3.80	2.60	15.60	0.20
Folio	7.59	0.70	1.80	70.23	0.50
GSM8K	16.40	11.00	2.30	56.50	0.10
HumanEVAL	84.15	37.20	53.66	84.76	50.00
LogicbenchMCQ	0.00	0.42	0.00	50.21	0.00
Logicbench Yes/No	0.40	0.80	0.20	73.60	0.20
MBPP	30.75	13.37	9.63	46.52	1.87
MultiRC	1.40	0.00	1.10	62.40	0.00
SST-2	13.50	5.70	4.40	19.30	8.10
SVAMP	31.71	14.71	5.43	61.00	0.29
WSC	11.85	0.15	1.52	22.34	0.00

Table 11: Percentage of Translations Removed After BLEU Score Filtering for **ENDIVE** Across Dialects and Datasets

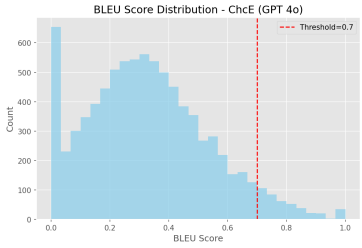
Dataset	AAVE (%)	ChcE (%)	CollSgE (%)	IndE (%)	JamE (%)
BoolQ	19.3	59.3	0.0	5.2	13.6
COPA	3.8	80.5	0.0	8.1	15.0
Folio	18.9	75.4	0.4	4.7	6.3
GSM8K	11.4	85.3	0.2	2.5	15.1
HumanEVAL	10.0	87.1	92.5	76.0	41.4
Logic Bench MCQ	16.2	78.4	1.0	2.1	18.8
Logic Bench Yes/No	12.6	68.1	0.6	4.4	12.1
MBPP	11.2	59.5	2.8	3.8	19.7
MultiRC	20.0	48.3	3.9	12.8	11.3
SST-2	15.2	47.1	4.0	8.7	13.7
SVAMP	21.4	60.2	1.3	7.2	14.6
WSC	18.3	50.3	2.7	6.1	8.9

Table 12: Percentage of Translations Removed After BLEU Score Filtering for Multi-VALUE Across Dialects and Datasets

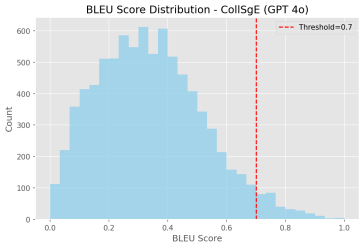
C BLEU Score Analysis Across Dialects for ENDIVE



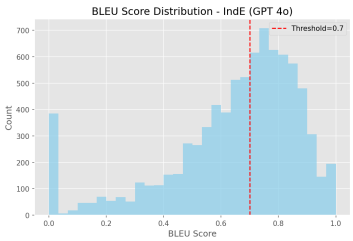
(a) AAVE



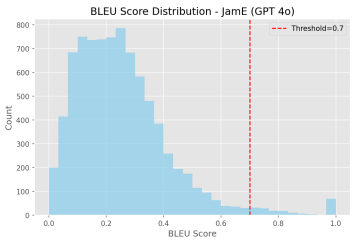
(b) ChcE



(c) CollSgE



(d) IndE



(e) JamE

Figure 1: BLEU score distributions for each dialect evaluated in this study. These histograms help visualize variance and score skewness.

C.1 Qualitative BLEU Analysis

Similar vs Threshold	SAE Premise	Dialect Premise	BLEU Score
Overly Similar (AAVE)	Matthew had 29 crackers and 30 cakes. If Matthew gave equal numbers of crackers and cakes to his 2 friends, how many cakes did each person eat?	Matthew had 29 crackers and 30 cakes. If he done gave equal numbers of crackers and cakes to his 2 friends, how many cakes each person eat?	0.9510699416
Overly Similar (ChcE)	Zachary did 46 push-ups and 58 crunches in gym class today. David did 38 more push-ups but 62 less crunches than Zachary.	Zachary did 46 push-ups and 58 crunches in gym class today. David did like 38 more push-ups but 62 less crunches than Zachary.	0.8127596564
Near Threshold (AAVE)	Helen the hippo and her friends are preparing for Thanksgiving at Helen's house. Helen baked 197 chocolate chip cookies and 46 raisin cookies yesterday. She also baked 75 raisin cookies and 66 chocolate chip cookies this morning.	Helen the hippo and her friends gettin' ready for Thanksgiving at Helen's crib. Helen done baked 197 chocolate chip cookies and 46 raisin cookies yesterday. And she baked 75 raisin cookies and 66 chocolate chip cookies this mornin'.	0.700020017
Near Threshold (CollSgE)	Jerry had 4 action figures and 22 books on a shelf in his room. Later he added 6 more action figures to the shelf.	Jerry got 4 action figures and 22 books on the shelf in his room. Later he add 6 more action figures to the shelf.	0.7094521095
Near Threshold (IndE)	Jack received 4 emails and sent 2 letters in the morning. He then received 6 emails and sent 8 letters in the afternoon.	Jack received 4 emails and sent 2 letters in the morning. Then in the afternoon, he received 6 emails and sent 8 letters.	0.7094521095

Table 13: Sample translations at various BLEU scores. Those above 0.70 are nearly identical to SAE, while those near 0.70 remain faithful yet show distinct dialect features.

D Metrics

Dataset	AAVE	IndE	JamE	CollSgE
BoolQ	0.6823 / **0.6881**	**0.7004** / 0.6927	0.6617 / **0.6648**	**0.6995** / 0.6915
COPA	**0.9864** / 0.9851	**0.9930** / 0.9908	**0.9876** / 0.9703	**0.9914** / 0.9911
MultiRC	**0.5623** / 0.5528	**0.7982** / 0.7728	**0.8151** / 0.4793	**0.6040** / 0.5753
SST-2	0.9588 / **0.9611**	**0.9711** / 0.9678	**0.9555** / 0.9412	**0.9721** / 0.9674
WSC	0.9074 / **0.9088**	**0.8986** / 0.4044	**0.7341** / 0.4013	**0.9121** / 0.9112
MBPP	0.7617 / **0.8188**	**0.9432** / 0.9162	0.6289 / **0.7370**	**0.9536** / 0.9347
GSM8K	**0.7201** / 0.7100	**0.7237** / 0.7230	0.6640 / **0.6778**	**0.7236** / 0.6961
SVAMP	**0.7923** / 0.7904	**0.8418** / 0.7632	**0.7896** / 0.5346	**0.7938** / 0.7638
FOLIO	**0.5797** / 0.5663	**0.5618** / 0.5536	0.5319 / **0.5391**	**0.6076** / 0.5464
Logic Bench MCQ	0.4953 / **0.7847**	**0.8841** / 0.7421	**0.7808** / 0.4541	**0.6751** / 0.4447
Logic Bench YN	**0.4742** / 0.2183	**0.8139** / 0.7401	0.4386 / **0.7788**	0.4331 / **0.6732**
Average	**0.7151** / 0.7063	**0.8087** / 0.7284	**0.6717** / 0.6003	**0.7359** / 0.6824

Table 14: *Lexical Diversity Scores across Dialects and Datasets (*ENDIVE*/Multi-VALUE).* Bold indicates the higher score.

Model	Dataset	IndE	AAVE	CollSgE	JamE
Gemini 1.5	BoolQ	**100.00** / 0.00	**100.00** / 0.00	**100.00** / 0.00	**100.00** / 0.00
	COPA	**87.56** / 12.44	**91.86** / 8.14	**70.02** / 29.98	**93.15** / 6.85
	MultiRC	**100.00** / 0.00	**100.00** / 0.00	**100.00** / 0.00	**100.00** / 0.00
	SST-2	**84.74** / 15.26	**93.96** / 6.04	**77.49** / 22.51	**94.46** / 5.54
	FOLIO	**96.58** / 3.42	**94.95** / 5.05	**95.70** / 4.30	**98.63** / 1.37
	HumanEval	**100.00** / 0.00	N/A	N/A	**100.00** / 0.00
	MBPP	**100.00** / 0.00	**100.00** / 0.00	**84.98** / 15.02	**99.40** / 0.60
	GSM8K	**99.00** / 1.00	**99.27** / 0.73	**99.78** / 0.22	**98.77** / 1.23
	SVAMP	**97.91** / 2.09	**99.73** / 0.27	**98.86** / 1.14	**94.39** / 5.61
	Logic Bench MCQ	**99.56** / 0.44	**100.00** / 0.00	**99.56** / 0.44	**100.00** / 0.00
	Logic Bench YN	**100.00** / 0.00	**100.00** / 0.00	**98.74** / 1.26	**99.76** / 0.24
	Average	**97.11** / 2.89	**97.99** / 2.01	**92.99** / 7.01	**97.88** / 2.12
GPT-4o	BoolQ	**99.24** / 0.76	**99.49** / 0.51	**99.73** / 0.27	**99.65** / 0.35
	COPA	**79.43** / 20.57	**92.39** / 7.61	**73.92** / 26.08	**93.79** / 6.21
	MultiRC	**100.00** / 0.00	**100.00** / 0.00	**100.00** / 0.00	**100.00** / 0.00
	SST-2	**80.61** / 19.39	**89.34** / 10.66	**87.75** / 12.25	**88.11** / 11.89
	FOLIO	**88.36** / 11.64	**94.91** / 5.09	**94.70** / 5.30	**91.75** / 8.25
	HumanEval	**100.00** / 0.00	N/A	N/A	**100.00** / 0.00
	MBPP	**99.48** / 0.52	**96.70** / 3.30	**91.59** / 8.41	**98.81** / 1.19
	GSM8K	**97.00** / 3.00	**94.88** / 5.12	**92.62** / 7.38	**91.01** / 8.99
	SVAMP	**97.49** / 2.51	**93.30** / 6.70	**88.62** / 11.38	**79.20** / 20.80
	Logic Bench MCQ	**95.13** / 4.87	**100.00** / 0.00	**92.81** / 7.19	**99.24** / 0.76
	Logic Bench YN	**93.60** / 6.40	**100.00** / 0.00	**94.56** / 5.44	**98.54** / 1.46
	Average	**93.78** / 6.22	**96.22** / 3.78	**91.72** / 8.28	**94.11** / 5.89

Table 15: *Preference Scores (%) for **EnDive vs. Multi-VALUE** using **Gemini 1.5** and **GPT-4o**. Each cell shows the percentage of responses where EnDive was preferred over Multi-VALUE.*

E LLM Dataset Evaluation Results

Dataset	AAVE		ChcE		CollSgE		IndE		JamE	
	CoT	*SAE CoT*	*CoT*	*SAE CoT*	*CoT*	*SAE CoT*	*CoT*	*SAE CoT*	*CoT*	*SAE CoT*
Language Understanding										
BoolQ	91.22	**92.92**	90.08	**92.97**	91.25	**93.07**	**93.34**	91.76	92.62	**93.29**
COPA	98.41	**99.31**	97.92	**99.11**	98.60	**99.42**	98.95	**99.12**	**99.53**	98.09
MultiRC	89.11	**91.37**	88.19	**91.22**	89.28	**91.43**	89.59	**91.61**	88.55	**91.33**
WSC	65.95	**90.11**	64.08	**90.33**	65.42	**90.18**	67.79	**90.47**	66.64	**90.38**
SST-2	93.35	**94.87**	92.69	**94.72**	93.02	**94.96**	93.54	**95.12**	92.28	**94.90**
Algorithmic Understanding										
HumanEval	97.09	**97.68**	96.82	**97.43**	96.94	**97.55**	97.37	**97.73**	97.19	**97.62**
MBPP	**95.17**	94.42	**94.53**	94.17	**94.89**	94.52	**95.32**	95.10	**95.10**	94.64
Math										
GSM8K	**96.34**	94.72	**95.82**	94.61	**96.13**	94.96	**96.83**	95.08	**96.48**	95.38
SVAMP	96.29	**96.97**	95.96	**96.82**	96.56	**97.20**	96.82	**97.27**	96.65	**97.18**
Logic										
FOLIO	75.94	**79.27**	75.07	**79.03**	76.28	**79.42**	77.06	**79.67**	76.62	**79.60**
Logic Bench MCQ	87.15	**89.32**	86.46	**89.17**	87.51	**89.52**	88.27	**89.82**	87.68	**89.73**
Logic Bench YN	80.65	**83.72**	86.80	**88.10**	82.42	**83.47**	81.80	**84.16**	80.98	**84.02**
Average	88.89	**92.06**	88.70	**92.31**	89.02	**92.14**	89.72	**92.24**	89.19	**92.18**

Table 16: Gemini 2.5 Pro accuracy (%) across 12 tasks, with CoT vs. SAE CoT and per-dialect averages (bolded for higher values).

Dataset	AAVE				ChcE				CollSgE				IndE				JamE			
	ZS	*CoT*	*SAE ZS*	*SAE CoT*	*ZS*	*CoT*	*SAE ZS*	*SAE CoT*	*ZS*	*CoT*	*SAE ZS*	*SAE CoT*	*ZS*	*CoT*	*SAE ZS*	*SAE CoT*	*ZS*	*CoT*	*SAE ZS*	*SAE CoT*
Language Understanding																				
BoolQ	88.31	87.68	90.43	**91.57**	87.63	88.44	90.25	**91.38**	88.25	88.04	90.84	**91.45**	88.25	86.47	90.61	**91.33**	88.04	87.61	90.72	**91.41**
COPA	**98.35**	98.32	97.22	97.85	97.92	**98.52**	97.47	98.02	97.54	**98.34**	97.18	97.95	**98.58**	98.33	97.64	98.20	96.39	**97.77**	97.11	97.73
MultiRC	88.24	89.54	89.02	**89.77**	88.30	87.37	89.09	**89.65**	89.28	88.72	89.11	**89.79**	86.70	88.74	89.15	**89.70**	87.70	89.15	89.21	**89.72**
WSC	72.13	71.54	81.67	**88.43**	55.10	54.45	81.52	**88.29**	68.36	78.24	81.75	**88.37**	60.23	63.12	81.49	**88.41**	61.33	67.18	81.57	**88.45**
SST-2	91.79	92.81	89.96	**93.14**	90.24	89.92	89.78	**93.02**	89.75	91.18	89.92	**93.20**	90.71	90.56	89.89	**93.07**	88.90	89.42	89.84	**93.11**
Algorithmic Understanding																				
HumanEval	88.46	**96.15**	94.12	93.87	97.09	**99.02**	94.31	93.76	**96.05**	91.89	94.22	93.91	**96.00**	95.83	94.07	93.85	91.46	92.68	**94.15**	93.97
MBPP	**88.42**	85.66	85.93	74.28	86.73	**86.88**	85.82	74.15	86.98	**87.13**	85.94	74.32	**86.00**	85.93	85.76	74.40	**88.49**	**88.49**	85.88	74.36
Math																				
GSM8K	74.46	66.29	89.45	**90.21**	52.76	66.29	89.14	**90.18**	40.74	64.38	89.36	**90.10**	82.70	66.67	89.23	**90.30**	67.92	66.27	89.41	**90.25**
SVAMP	92.68	69.33	94.10	**94.52**	68.01	73.53	94.07	**94.43**	62.03	70.24	94.21	**94.55**	**94.42**	70.96	94.12	94.47	93.45	70.01	94.18	**94.49**
Logic																				
FOLIO	61.19	63.24	73.89	**74.51**	61.97	62.64	73.58	**74.67**	64.39	66.46	73.42	**74.83**	69.13	63.76	73.74	**74.55**	63.65	65.69	73.69	**74.47**
LogicBench MCQ	**84.73**	72.42	82.55	83.64	**83.86**	72.21	82.42	83.79	**84.34**	72.33	82.61	83.52	83.66	68.07	82.49	**83.71**	**85.69**	72.33	82.67	83.68
LogicBench Y/N	68.45	75.91	75.62	**76.94**	67.33	76.55	75.49	**76.81**	66.49	75.94	75.74	**76.88**	70.15	76.30	75.53	**76.93**	67.19	76.49	75.67	**76.79**
Average	78.15	79.78	80.04	**83.10**	77.96	81.15	87.28	**88.78**	77.96	81.15	87.43	**88.83**	84.06	79.61	87.38	**88.82**	81.65	80.18	87.41	**88.79**

Table 17: Claude 3.5 Sonnet accuracy (%). The single highest value among ZS, CoT, SAE ZS, and SAE CoT in each dialect and dataset is shown in **bold**.

Dataset	AAVE				ChcE				CollSgE				IndE				JamE			
	ZS	*CoT*	*SAE ZS*	*SAE CoT*	*ZS*	*CoT*	*SAE ZS*	*SAE CoT*	*ZS*	*CoT*	*SAE ZS*	*SAE CoT*	*ZS*	*CoT*	*SAE ZS*	*SAE CoT*	*ZS*	*CoT*	*SAE ZS*	*SAE CoT*
Language Understanding																				
BoolQ	89.09	88.33	91.10	**91.75**	88.83	88.23	90.25	**91.10**	88.36	88.05	**91.50**	90.95	89.25	88.50	90.80	**91.30**	89.15	88.34	90.95	**91.20**
COPA	**97.87**	97.64	96.80	97.40	98.34	**98.54**	97.10	97.75	97.13	97.13	96.90	**97.45**	97.87	**98.34**	97.20	97.85	96.39	96.59	97.15	**97.60**
MultiRC	86.71	87.32	88.93	**89.76**	86.80	86.60	88.85	**89.65**	87.26	87.06	88.95	**89.75**	85.11	85.11	88.80	**89.60**	87.70	88.03	88.95	**89.83**
WSC	58.97	60.52	80.97	**88.55**	57.63	54.95	80.80	**88.40**	58.80	58.02	80.95	**88.53**	67.84	69.59	80.85	**88.35**	55.63	56.87	80.75	**88.45**
SST-2	90.17	90.29	89.88	**93.19**	89.61	89.08	89.85	**93.00**	89.23	89.02	89.75	**93.26**	89.71	88.85	89.90	**93.05**	87.92	86.72	89.95	**93.15**
Algorithmic Understanding																				
HumanEVAL	88.46	84.62	**94.00**	93.50	97.09	**99.03**	94.10	93.80	**97.37**	96.05	94.20	93.90	**100.00**	96.28	94.05	93.85	**100.00**	97.56	94.15	93.95
MBPP	84.56	83.92	**85.00**	73.81	81.00	79.00	**84.90**	74.00	82.54	**84.95**	84.02	73.85	81.00	79.00	**84.85**	74.10	83.92	83.92	**84.75**	74.05
Math																				
GSM8K	57.32	85.64	89.30	**90.15**	57.43	76.63	89.00	**90.25**	58.65	83.01	89.40	**90.50**	51.18	87.47	89.60	**90.10**	54.98	84.76	89.20	**90.71**
SVAMP	90.82	92.74	94.15	**94.59**	91.48	92.92	94.00	**94.40**	90.86	93.99	94.22	**94.62**	91.27	93.73	94.05	**94.55**	91.44	94.33	94.15	**94.65**
Logic																				
FOLIO	64.90	64.97	73.50	**74.90**	64.08	64.39	73.75	**75.30**	65.31	65.51	72.90	**74.45**	68.79	69.80	74.10	**75.00**	66.67	64.36	73.80	**75.10**
Logic Bench MCQ	79.05	78.95	82.65	**83.75**	78.31	62.47	82.40	**83.50**	79.71	77.57	82.84	**83.65**	75.94	70.00	82.30	**83.45**	78.41	76.63	82.59	**83.55**
Logic Bench YN	72.55	71.43	75.81	**76.95**	73.44	72.58	75.90	**77.00**	70.78	69.72	75.76	**76.85**	71.43	72.96	75.60	**76.90**	72.13	72.27	75.85	**77.05**
Average	80.04	82.20	86.84	**87.36**	80.34	80.37	86.74	**87.35**	80.50	82.43	86.86	**87.31**	80.78	83.30	86.84	**87.34**	80.36	82.53	86.85	**87.44**

Table 18: GPT-4o accuracy (%). The single highest value among ZS, CoT, SAE ZS, and SAE CoT in each dialect and dataset is shown in **bold**.

Dataset	AAVE		ChcE		CollSgE		IndE		JamE	
	CoT	SAE CoT	CoT	SAE CoT	CoT	SAE CoT	CoT	SAE CoT	CoT	SAE CoT
Language Understanding										
BoolQ	91.65	**92.03**	89.22	**91.35**	89.50	**92.20**	91.10	**92.10**	90.60	**92.05**
COPA	98.30	**98.45**	96.90	**98.20**	97.60	**98.40**	97.85	**98.35**	97.20	**98.25**
MultiRC	88.40	**91.00**	87.50	**91.05**	88.00	**91.00**	89.00	**91.20**	88.50	**90.75**
WSC	65.50	**89.50**	63.80	**89.65**	64.50	**89.55**	67.50	**89.70**	66.00	**89.60**
SST-2	93.50	**95.00**	92.20	**94.80**	92.90	**94.80**	93.10	**95.10**	91.30	**94.00**
Algorithmic Understanding										
HumanEval	96.50	**97.00**	96.00	**97.10**	96.20	**97.20**	96.20	**97.25**	96.00	**97.10**
MBPP	**94.00**	93.00	**93.50**	93.40	**93.20**	93.10	**93.80**	93.60	93.20	**93.30**
Math										
GSM8K	**95.00**	93.50	**94.50**	93.80	**95.10**	94.00	**95.50**	94.40	**94.50**	94.10
SVAMP	95.80	**96.00**	95.30	**95.90**	95.50	**96.10**	95.70	**96.30**	95.40	**96.20**
Logic										
FOLIO	65.50	**89.50**	63.80	**89.65**	64.50	**89.55**	67.50	**89.70**	66.00	**89.60**
LogicMCQ	87.15	**89.32**	86.46	**89.17**	87.51	**89.52**	88.27	**89.82**	87.68	**89.73**
LogicYN	80.65	**83.72**	86.80	**88.10**	82.42	**83.47**	81.80	**84.16**	80.98	**84.02**
Average	89.13	**93.15**	88.54	**93.39**	89.14	**93.50**	90.34	**94.07**	89.40	**93.14**

Table 19: o1 accuracy (%). **Bold** marks the higher score within each dataset row and within the averages.

Dataset	AAVE				ChcE				CollSgE				IndE				JamE			
	ZS	*CoT*	*SAE ZS*	*SAE CoT*	*ZS*	*CoT*	*SAE ZS*	*SAE CoT*	*ZS*	*CoT*	*SAE ZS*	*SAE CoT*	*ZS*	*CoT*	*SAE ZS*	*SAE CoT*	*ZS*	*CoT*	*SAE ZS*	*SAE CoT*
Language Understanding																				
BoolQ	90.29	90.05	91.47	**91.92**	89.74	89.89	91.25	**91.61**	89.89	89.79	91.53	**91.78**	90.75	90.50	91.62	**91.95**	89.65	89.45	91.58	**91.83**
COPA	97.16	96.93	96.77	**97.42**	96.88	96.47	97.20	**97.45**	97.33	97.33	97.10	**97.40**	**98.10**	**98.10**	97.36	97.81	94.59	94.99	97.01	**97.37**
MultiRC	86.92	86.41	89.07	**89.76**	86.50	87.10	89.13	**89.67**	87.26	86.75	89.10	**89.79**	86.44	85.11	89.15	**89.71**	87.20	87.10	89.20	**89.73**
WSC	54.83	51.55	81.69	**88.42**	54.95	50.53	81.55	**88.29**	54.71	51.54	81.71	**88.39**	62.57	53.82	81.49	**88.41**	54.23	53.19	81.61	**88.47**
SST-2	91.91	92.25	89.97	**93.12**	91.62	91.30	89.80	**93.04**	90.06	89.64	89.94	**93.19**	91.08	90.95	89.86	**93.08**	89.55	89.01	89.82	**93.10**
Algorithmic Understanding																				
HumanEVAL	92.31	92.31	**94.10**	93.85	**97.09**	96.12	94.32	93.78	92.11	**96.05**	94.20	93.91	**96.00**	**96.00**	94.05	93.85	91.46	91.46	**94.14**	93.96
MBPP	85.29	**86.49**	85.92	74.31	**86.73**	85.80	85.84	74.17	**86.98**	85.50	85.95	74.35	84.00	83.00	**85.79**	74.42	**86.92**	**86.92**	85.86	74.38
Math																				
GSM8K	60.86	84.05	89.54	**90.27**	59.54	77.17	89.25	**90.10**	51.28	78.40	89.38	**90.19**	60.36	87.13	89.41	**90.32**	60.07	80.86	89.29	**90.22**
SVAMP	92.68	90.99	94.11	**94.51**	92.77	91.96	94.05	**94.40**	92.46	90.63	94.22	**94.54**	92.77	91.58	94.09	**94.48**	92.99	90.11	94.18	**94.47**
Logic																				
FOLIO	62.27	63.57	73.61	**74.15**	63.68	62.88	73.80	**74.20**	65.62	65.21	73.91	**74.43**	68.12	68.12	73.74	**74.57**	65.56	65.16	73.83	**74.49**
LogicBench MCQ	78.41	73.96	82.52	**83.65**	79.58	73.85	82.48	**83.70**	80.38	73.54	82.60	**83.57**	79.83	74.48	82.50	**83.74**	78.87	72.92	82.66	**83.71**
LogicBench Y/N	**77.45**	76.12	75.63	76.97	76.69	75.56	75.51	**76.83**	**77.44**	75.40	75.74	76.92	**78.06**	76.02	75.55	76.91	**77.21**	75.69	75.66	76.78
Average	80.86	82.06	87.03	**87.36**	81.31	81.55	87.02	**87.27**	80.46	81.65	87.12	**87.37**	82.34	82.90	87.05	**87.44**	80.69	81.40	87.07	**87.38**

Table 20: DeepSeek-v3 accuracy (%). The single highest value among ZS, CoT, SAE ZS, and SAE CoT in each dialect and dataset is shown in **bold**.

Dataset	AAVE				ChcE				CollSgE				IndE				JamE			
	ZS	*CoT*	*SAE ZS*	*SAE CoT*	*ZS*	*CoT*	*SAE ZS*	*SAE CoT*	*ZS*	*CoT*	*SAE ZS*	*SAE CoT*	*ZS*	*CoT*	*SAE ZS*	*SAE CoT*	*ZS*	*CoT*	*SAE ZS*	*SAE CoT*
Language Understanding																				
BoolQ	86.70	87.13	88.42	**89.10**	85.21	86.32	88.15	**89.05**	86.21	85.60	88.31	**89.14**	86.25	86.50	88.23	**89.09**	84.92	86.83	88.28	**89.12**
COPA	95.98	**96.45**	94.78	95.43	94.59	**95.84**	94.63	95.38	94.66	**95.48**	94.57	95.29	94.79	**95.26**	94.81	95.32	93.39	94.79	94.74	**95.22**
MultiRC	84.08	84.48	88.15	**88.75**	82.90	83.70	88.12	**88.63**	84.63	85.44	88.08	**88.79**	82.71	83.51	88.17	**88.70**	85.00	84.60	88.21	**88.72**
WSC	54.31	53.62	79.68	**85.42**	55.93	49.77	79.54	**85.29**	54.63	53.86	79.71	**85.38**	54.39	55.56	79.51	**85.41**	53.35	50.70	79.63	**85.45**
SST-2	90.64	**91.91**	89.72	92.88	90.35	**90.77**	89.58	92.80	87.34	**89.54**	89.76	92.97	89.34	**89.84**	89.69	92.85	87.16	**88.14**	89.64	92.89
Algorithmic Understanding																				
HumanEVAL	**100.00**	**100.00**	93.94	93.78	**100.00**	99.03	94.13	93.65	**100.00**	98.68	94.21	93.89	**100.00**	**100.00**	94.07	93.83	**100.00**	98.78	94.12	93.91
MBPP	74.14	**80.69**	83.12	80.31	79.32	**80.25**	83.01	74.09	82.84	**85.50**	83.23	74.17	76.00	**78.50**	82.97	74.23	76.02	**78.20**	83.05	74.21
Math																				
GSM8K	56.34	**75.84**	58.40	58.30	54.72	**75.39**	58.40	58.30	55.17	**76.25**	58.40	58.30	57.93	**77.47**	58.40	58.30	52.75	**72.47**	58.40	58.30
SVAMP	74.27	**77.82**	77.14	74.43	77.05	**75.71**	77.14	74.43	73.26	**77.64**	77.14	74.43	79.85	**75.09**	77.14	74.43	73.07	**78.65**	77.14	74.43
Logic																				
FOLIO	51.03	41.73	**52.25**	52.15	54.02	41.15	**52.25**	52.15	53.20	40.79	**52.25**	52.15	51.68	43.62	**52.25**	52.15	51.61	42.57	**52.25**	52.15
LogicBench MCQ	60.62	40.92	**67.50**	66.67	62.55	38.57	**67.50**	66.67	61.25	41.75	**67.50**	66.67	61.09	39.08	**67.50**	66.67	59.38	39.46	**67.50**	66.67
LogicBench Y/N	61.04	**63.82**	62.83	61.97	63.48	**66.67**	62.83	61.97	60.95	**63.92**	62.83	61.97	61.48	**70.92**	62.83	61.97	61.73	**64.23**	62.83	61.97
Average	74.10	74.53	77.99	**78.27**	75.01	75.01	77.94	**77.70**	74.51	80.59	86.14	**86.61**	74.29	74.26	84.53	**86.63**	76.63	80.56	86.11	**86.60**

Table 21: GPT-4o-mini accuracy (%). The single highest value among ZS, CoT, SAE ZS, and SAE CoT in each dialect and dataset is shown in **bold**.

Dataset	AAVE				ChcE				CollSgE				IndE				JamE			
	ZS	*CoT*	*SAE ZS*	*SAE CoT*	*ZS*	*CoT*	*SAE ZS*	*SAE CoT*	*ZS*	*CoT*	*SAE ZS*	*SAE CoT*	*ZS*	*CoT*	*SAE ZS*	*SAE CoT*	*ZS*	*CoT*	*SAE ZS*	*SAE CoT*
LANGUAGE UNDERSTANDING																				
BoolQ	81.39	79.49	83.49	**85.29**	82.36	83.07	**86.93**	85.71	79.22	77.34	81.09	**84.24**	75.27	74.06	78.86	**80.45**	77.20	77.56	**81.44**	80.23
COPA	98.06	98.85	99.36	**100.00**	96.29	97.17	98.15	**98.77**	97.13	97.95	98.67	**99.31**	95.42	96.28	97.30	**98.29**	96.15	97.03	98.02	**98.76**
MultiRC	87.05	85.23	86.18	**88.11**	86.80	84.92	87.56	**89.46**	88.03	86.71	88.92	**90.25**	85.11	83.88	86.15	**88.09**	87.70	86.53	89.01	**90.22**
WSC	61.17	59.88	63.12	**65.04**	57.63	55.49	59.82	**61.64**	58.80	57.02	60.33	**62.11**	67.84	66.31	69.02	**70.19**	55.63	56.87	59.45	**60.88**
SST-2	90.64	91.91	92.82	**93.82**	90.35	90.77	91.89	**92.76**	89.23	89.05	90.47	**92.31**	89.71	89.84	91.04	**92.85**	87.16	88.14	90.56	**92.89**
ALGORITHMIC UNDERSTANDING																				
HumanEVAL	85.67	86.12	88.45	**90.12**	93.04	93.51	95.03	**96.87**	94.22	94.68	96.45	**97.30**	95.83	96.28	97.14	**98.07**	91.46	92.68	94.14	**95.02**
MBPP	78.09	77.02	81.14	**83.27**	83.12	82.54	85.33	**87.09**	82.54	82.01	85.14	**86.97**	81.00	80.71	84.09	**86.03**	82.84	82.45	85.04	**87.36**
MATH																				
GSM8K	71.39	71.59	73.20	**73.61**	65.21	65.57	67.13	**69.09**	58.65	58.41	60.18	**62.09**	82.70	83.26	84.55	**87.61**	67.92	68.01	69.78	**71.54**
SVAMP	81.06	81.59	89.54	**90.21**	91.48	92.92	94.00	**94.40**	90.86	93.99	94.22	**94.62**	91.27	93.73	94.05	**94.55**	91.44	94.33	94.15	**94.65**
LOGIC																				
FOLIO	70.94	71.63	73.77	**75.04**	68.25	67.14	70.90	**72.51**	65.31	65.51	72.90	**74.45**	69.13	69.80	74.10	**75.00**	63.65	65.69	73.80	**75.10**
LogicBench MCQ	86.50	88.20	90.04	**91.37**	78.31	82.47	86.40	**87.55**	79.71	82.15	85.66	**87.22**	75.94	81.37	85.12	**87.05**	78.41	83.14	86.82	**88.73**
LogicBench Y/N	80.20	79.55	82.30	**83.76**	73.44	72.58	75.42	**77.09**	70.78	69.72	75.76	**76.85**	71.43	72.96	75.60	**76.90**	72.13	72.27	75.85	**77.05**
Average	82.59	82.69	85.66	**87.49**	78.54	78.08	81.12	**82.94**	78.43	78.41	81.07	**83.00**	81.09	81.52	84.07	**86.12**	79.02	79.14	82.07	**83.20**

Table 22: LLaMa-3-8B Instruct accuracy (%). **Bold** indicates superior performance within dialect pairs and average.

F Qualitative Analysis

Rubric Item	Multi-VALUE	ENDIVE
Consistent past-tense forms	13 campers *goed* rowing and 59 campers *goed* hiking in the morning. 21 campers *goed* rowing in the afternoon.	**So like**, 13 campers went rowing and 59 campers went hiking in the morning, **you know**? And then in the afternoon, 21 campers went rowing.
ChcE auxiliaries / progressive "be"	James write a 3-page letter to 2 different friend twice a week. How many pages do write a year?	James **be writin'** a 3-page letter to 2 different **homies** twice a week. How many pages he **be writin'** in a year?
Subject-verb agreement ("does … got")	There is 5 houses on a street, and each of the first four houses have 3 gnomes in the garden. If there is 20 gnomes in total on the street, how many gnomes do the fifth house have?	… how many gnomes **does** the fifth house **got**?
Conversational flow + plurals	Joy might can read 8 page of a book in 20 minute. How many hours might will it take her to read 120 page?	Joy can read **like** 8 pages … **So like** how many hours it's **gonna** take her to read 120 pages?
Discourse framing with "only/like"	Jake have 5 fewer peaches than Steven. Steven have 18 more peaches than Jill.	**So check it out**, Jake **got like** 5 less peaches than Steven, **right**? …

Table 23: ChcE examples. **BrickRed** = core ChcE morpho-syntactic markers (habitual/progressive *be*, "got like," local nouns such as *homies*). **MidnightBlue** = discourse fillers / stance markers ("so like," "you know," "right").

Rubric Item	Multi-VALUE	ENDIVE
Definite/indefinite article use	Vic DiCara plays guitar and bass. A only style of musics Vic plays it are punk musics.	The **only** style of music that Vic DiCara is playing is punk music.
"Only" as focus particle; IndE progressive	All eels are fishs. No fishs are plants. Everything have displayed collection is either plant or animal.	All eels are fish **only**. No fish **are being** plants. Everything shown in the collection is either a plant or an animal.
Consistent verb tenses	If legislator is found it guilty stealing governments funds, it would be suspended office.	If a legislator is found guilty of stealing government funds, they would be suspended from office.
Subscription example with IndE verb choice	All customers James' family is subscribing AMC A-List are like eligible to watch three movie every week any additional fees.	James' family **subscribes** to AMC A-List or HBO services. Customers who prefer TV series will not watch TV series in cinemas.
Code-switching with rupees / paise	Peter goes store to buy sodas. sodas cost $0.25 ounce. had brought $2 him and leaves $0.50. How many ounce sodas buy?	Peter goes to the shop to buy a cold drink. The cold drink costs 25 **paise** an ounce. He brought 2 **rupees** with him and leaves with 50 **paise**. How many ounces of cold drink did he buy?

Table 24: IndE examples. **BrickRed** = IndE grammatical features (focus particle *only*, progressive "are being," local verb choices like *subscribes*). **MidnightBlue** = lexical code-switches to Indian currency terms (*rupees, paise*).

Rubric Item	Multi-VALUE	ENDIVE
Sentence-final particles "lah/ah"	All social medium application containing chat feature software.	All the social media apps with chat features **ah**, all software one **lah**.
Auxiliary omission + "kena"	Any convicted criminal that like innocent is not like truly guilty.	Any convicted criminal who **kena** innocent one, not really guilty **lah**.
"Kena" for passive/adversity	Everyone convicted murders goes prison.	Anyone **kena** convicted of murder **sure** go prison **one**.
Discourse fillers + final "one"	Roy Richardson one was cricketer who play Sint Maarten, constituent country.	Roy Richardson **ah**, he was a cricketer who play for Sint Maarten, **you know**, that place part of another country **one**.
Other particles "siah/lor/leh"	UFC Fight Night, Sadollah have been scheduled fight Musoke.	Sadollah fight Akiyama at UFC Fight Night, **siah**.

Table 25: CollSgE examples. **BrickRed** = core Singlish morpho-syntactic elements (passive *kena*, stance/focus particles *lah, one, siah*). **MidnightBlue** = conversational fillers / sentence-medial particles (*ah*, "sure," "you know").

G Failure Mode Analysis

To better understand where models struggle with dialectal variation, we conduct a detailed error analysis across all seven evaluated models: **Gemini 2.5 Pro** (DeepMind, 2025), **o1** (OpenAI, 2024b), **Claude 3.5 Sonnet** (Anthropic, 2024), **GPT-4o** (OpenAI, 2024a), **GPT-4o-mini** (OpenAI, 2024a), **DeepSeek-v3** (DeepSeek-AI, 2024), and **LLaMa-3-8B Instruct** (META, 2024). We focus on examples where models answered correctly in SAE but failed when presented with equivalent prompts rewritten in an underrepresented dialect.

G.1 Concrete Failure Examples

Table 26 provides representative cases from BoolQ, GSM8K, and FOLIO where dialectal phrasing alone led to failure. Notably, even high-performing models like **o1** and **Claude 3.5 Sonnet** falter under syntactic and stylistic shifts.

Task	SAE Input (Correct)	Dialect Input (Incorrect)	Model(s)
BoolQ	"Did the committee approve the bill?"	"The committee done approved the bill, right?"	Claude 3.5, o1
GSM8K	"If he had 12 pencils and gave away 4, how many are left?"	"He got 12 pencils. Gave 4 away. What's he got now?"	o1, GPT-4o-mini
FOLIO	"He had planned the party, but no one came."	"He been had that party planned, ain't nobody show up."	GPT-4o, LLaMa-3-8B

Table 26: Examples where models answered correctly on SAE inputs but failed on equivalent dialectal rewrites.

G.2 Observed Error Patterns

Several consistent trends emerged:

- **Grammar Sensitivity:** Non-SAE syntactic constructions like "done approved" or "been had" caused errors in BoolQ and FOLIO, even for top models like **o1** and **Claude 3.5**. These structures led to misparsing or incorrect entailment.
- **Math Disruption:** Informal arithmetic phrasing in GSM8K—such as omitted verbs or compressed structure—disrupted models like **o1**, **GPT-4o-mini**, and **DeepSeek-v3**, all of which performed well on SAE inputs but stumbled with dialect rewrites.
- **Template Overfitting: LLaMa-3-8B** showed high sensitivity to phrasing shifts across nearly all tasks, often relying heavily on seen formats. **GPT-4o-mini** similarly failed when inputs diverged from familiar prompt templates, especially in logic and commonsense reasoning.
- **Coreference Confusion: GPT-4o** frequently misinterpreted coreferential statements in dialects like AAVE and JamE, despite handling the same inputs well in SAE. Its performance drop was most pronounced in discourse-heavy datasets like FOLIO and LogicBench.

G.3 Comparative Robustness

These findings highlight that even models with high overall accuracy—such as **o1** (OpenAI, 2024b) and **Gemini 2.5 Pro** (DeepMind, 2025)—are not immune to systematic failures when exposed to syntactic or pragmatic variation in underrepresented dialects. Despite strong aggregate metrics, performance can degrade sharply on inputs that deviate from SAE. For instance, **Claude 3.5 Sonnet** (Anthropic, 2024), while strong on algorithmic tasks, faltered on informal sentence structures. **GPT-4o** (OpenAI, 2024a) also showed sensitivity to dialect-specific phrasing, particularly in logic and reading comprehension. More compact models like **GPT-4o-mini** and **LLaMa-3-8B** (META, 2024) were especially brittle, struggling with even minor rephrasings. These patterns underscore a broader concern: conventional evaluations can obscure failure modes that disproportionately affect speakers of non-standard dialects. Dialect-aware evaluation is essential for ensuring equitable, reliable model behavior across diverse English varieties.

H Translation Prompts

Here are examples of African American Vernacular English (AAVE):
1. I was bewildered, but I knew dat it was no gud asking his ass to explain.
2. Cochran pontificated windily for da camera.
3. I don't want them to follow in my footsteps, as I ain't go to no college, but I want them to go.

Here is the input text: {text}
Please rewrite the input text in African American Vernacular English (AAVE).

Table 27: Few-Shot Prompt for Translating SAE to AAVE

Here are examples of Chicano English (ChcE):
1. When people wanna fight me I'm like "well okay, well then I'll fight you."
2. They were saying that they had a lot of problems at Garner because it was a lot of fights and stuff.
3. I ain't really thinking about getting with J. or any other guy.

Here is the input text: {text}
Please rewrite the input text in Chicano English (ChcE).

Table 28: Few-Shot Prompt for Translating SAE to ChcE

Here are examples of Colloquial Singapore English (Singlish) (CollSgE):
1. But after a while it become quite senseless to me.
2. And got to know this kind-hearted scholar who shelter her with Ø umbrella when it was raining.
3. The cake John buy one always very nice to eat.

Here is the input text: {text}
Please rewrite the input text in Colloquial Singapore English (Singlish) (CollSgE).

Table 29: Few-Shot Prompt for Translating SAE to CollSgE

Here are examples of Indian English (IndE):
1. It was not too much common. Getting the accommodation has become very much difficult.
2. During monsoon we get lot of rain and then gets very soggy and sultry.
3. This is the second time that such an object had been sighted here.

Here is the input text: {text}
Please rewrite the input text in Indian English (IndE).

Table 30: Few-Shot Prompt for Translating SAE to IndE

Here are examples of Jamaican English (JamE):
1. Hill had initially been indicted with the Canute and the Michelle Saddler and their three companies.
2. The autopsy performed on Mae's torso shortly after it was found, revealed that her body was cut into pieces by a power machine saw.
3. The culture of the region has been unique in combining British and Western influences with African and Asian lifestyles.

Here is the input text: {text}
Please rewrite the input text in Jamaican English (JamE).

Table 31: Few-Shot Prompt for Translating SAE to JamE

I Evaluation Prompts

Given a mathematics problem, determine the answer. Simplify your answer as much as possible and encode the final answer in <answer></answer> (e.g., <answer>42</answer>).
Context: {problem}
Question: {question}
Answer:
If CoT: Let's think about this step by step before finalizing the answer.

Table 32: Prompt for SVAMP Evaluation

Given a coding problem, produce a Python function that solves the problem. Provide your entire code in <answer></answer> (e.g., <answer>def solve(): pass</answer>).
Problem: {problem}
Test Cases: {test_cases}
Answer:
If CoT: Let's think step by step about the problem-solving process before coding.

Table 33: Prompt for MBPP Evaluation

Given a yes/no question, answer yes or no. Provide your final answer in <answer></answer> (e.g., <answer>yes</answer>).
Context: {context}
Question: {question}
Answer:
If CoT: Let's think step by step before arriving at the answer.

Table 34: Prompt for LogicBenchYN Evaluation

Given a multiple-choice question with 4 choices, pick the correct choice number (1, 2, 3, or 4). Provide your final answer in <answer></answer> (e.g., <answer>2</answer>).
Context: {context}
Choices:
1) {choice1}
2) {choice2}
3) {choice3}
4) {choice4}
Answer:
If CoT: Let's analyze each choice step by step before determining the correct one.

Table 35: Prompt for LogicBenchMCQ Evaluation

Given a coding problem, produce a Python function that solves the problem. Provide your entire code in <answer></answer> (e.g., <answer>def solve(): pass</answer>).
Problem: {prompt_text}
Test Cases: {test_cases}
Answer:
If CoT: Let's break the problem down step by step before writing the code.

Table 36: Prompt for HumanEVAL Evaluation

Given a mathematics problem, determine the answer. Simplify your answer as much as possible and encode the final answer in <answer></answer> (e.g., <answer>1</answer>).
Problem: {problem}
Answer:
If CoT: Let's carefully solve the problem step by step before arriving at the final numeric answer.

Table 37: Prompt for GSM8K Evaluation

Given premises and a conclusion, determine whether the conclusion is True, False, or Uncertain. Provide your final answer in <answer></answer> (e.g., <answer>True</answer>).
Premises: {premises}
Conclusion: {conclusion}
Answer:
If CoT: Let's evaluate the premises step by step before deciding the conclusion.

Table 38: Prompt for FOLIO Evaluation

Given a pronoun resolution problem, determine whether Span 2 refers to Span 1. Provide your final answer in <answer></answer> (e.g., <answer>1</answer> for same or <answer>0</answer> for different).
Paragraph: {paragraph}
Span 1: {span1}
Span 2: {span2}
Answer:
If CoT: Let's analyze the relationship between Span 1 and Span 2 step by step before answering.

Table 39: Prompt for WSC Evaluation

Given a sentence, determine its sentiment. Provide your final answer in <answer></answer> (e.g., <answer>1</answer> for positive or <answer>0</answer> for negative).
Sentence: {sentence}
Answer:
If CoT: Let's analyze the sentiment of the sentence step by step before concluding.

Table 40: Prompt for SST-2 Evaluation

Given a paragraph, a question, and an answer choice, determine if the answer choice is correct. Provide your final answer in <answer></answer> (e.g., <answer>1</answer> for correct or <answer>0</answer> for incorrect).
Paragraph: {paragraph}
Question: {question}
Answer Choice: {answer_choice}
Answer:
If CoT: Let's analyze the paragraph and question step by step before confirming the correctness of the answer choice.

Table 41: Prompt for MultiRC Evaluation

Given a premise and two choices, pick which choice is more plausible. Provide your final answer in <answer></answer> (e.g., <answer>0</answer> for the first choice or <answer>1</answer> for the second).
Premise: {premise}
Choice 1: {choice1}
Choice 2: {choice2}
Answer:
If CoT: Let's compare the plausibility of both choices step by step before finalizing.

Table 42: Prompt for COPA Evaluation

Given a passage and a yes/no question, label it as TRUE or FALSE. Provide your final answer in <answer></answer> (e.g., <answer>TRUE</answer>).
Passage: {passage}
Question: {question}
Answer:
If CoT: Let's carefully consider the passage and the question step by step before labeling the answer.

Table 43: Prompt for BoolQ Evaluation

J Fluency Scoring Prompt

```
You are an expert linguist capable of detailed chain-of-thought reasoning.
You are given two pieces of text:
1) Original Text (SAE) - the standard American English version.
2) Dialect Text - a translated or adapted version in the {dialect} dialect.
Please evaluate the Dialect Text for:
1) Fluency in {dialect}:
  - Grammar, syntax, word choice, and overall naturalness in {dialect}.
  - Consistency, flow, and readability in {dialect}.
2) Meaning Preservation:
  - Does the Dialect Text retain the same meaning or intent as the Original Text (SAE)?
  - Are there changes or omissions that alter the meaning?
Use the following 1-7 scoring rubric (focused on fluency, but keep meaning in mind):
- 1: Completely unnatural, pervasive errors, nearly unintelligible.
- 2: Major issues in accuracy/naturalness, very awkward for {dialect}.
- 3: Noticeable errors or unnatural phrasing, partial alignment with {dialect}.
- 4: Average fluency, some issues; mostly understandable in {dialect}.
- 5: Good fluency, minor errors; consistent with {dialect}.
- 6: Very good fluency, rare issues; flows smoothly in {dialect}.
- 7: Excellent fluency, fully natural, error-free, perfectly aligned with {dialect}.
Instructions:
1. Provide a chain-of-thought explanation comparing meaning and evaluating fluency.
2. End with a single line: "Fluency Score: X" (where X is an integer 1-7).
Begin your detailed chain-of-thought analysis now.
```

Table 44: Prompt for Fluency Evaluation

K Preference Tests Prompt

```
You are an expert linguist with a strong understanding of {dialect}.
You are given:
1) Original Text (SAE) – a standard American English version for reference.
2) Translation A – a version in the {dialect} dialect.
3) Translation B – another version in the {dialect} dialect.
Your task: Decide which translation is better in the context of the {dialect}
dialect with respect to:
- Fluency (grammar, syntax, word choice, overall naturalness in {dialect})
- Accuracy (faithfulness to the original meaning, but expressed naturally
in {dialect})
- Readability (cohesion, clarity, and flow in {dialect})
- Cultural appropriateness (if relevant to {dialect})
Provide a detailed chain-of-thought (reasoning) as to how you weigh these
factors.
Then conclude with one final line in the exact format:
"Final preference score: X"
(where X = 1 if you prefer Translation A, or X = 2 if you prefer Translation
B).
Make sure you reveal your full thought process, then end with:
Final preference score: X
```

Table 45: Prompt for Translation Comparison Evaluation