Perceptions of Language Technology Failures from South Asian English Speakers

Faye Holt * **Will Held** * **Diyi Yang** †
*Georgia Institute of Technology, †Stanford University
{mholt9, wheld3}@gatech.edu
diyiy@stanford.edu

Abstract

English NLP systems have empirically worse performance for dialects other than Standard American English (SAmE). However, how these discrepancies impact use of language technology by speakers of non-SAmE global Englishes is not well understood. We focus on reducing this gap for South Asian Englishes (SAsE), a macro-group of regional varieties with cumulatively more speakers than SAmE, by surveying SAsE speakers about their interactions with language technology and compare their responses to a control survey of SAmE speakers. SAsE speakers are more likely to recall failures with language technology and more likely to reference specific issues with written language technology than their SAmE counterparts. Furthermore, SAsE speakers indicate that they modify both their lexicon and syntax to make technology work better, but that lexical issues are perceived as the most salient challenge. We then assess whether these issues are pervasive in more recently developed Large Language Models (LLMs), introducing two benchmarks for broader SAsE Lexical and Indian English Syntactic understanding and evaluating 11 families of LLMs on them.[1]

1 Introduction

Previous studies in Natural Language Processing have identified performance disparities between Standard American English (SAmE) and other English dialects (Blevins et al., 2016; Jørgensen et al., 2016a; Blodgett et al., 2016; Jurgens et al., 2017; Ziems et al., 2022a, 2023; Shan et al., 2023). However, the degree to which these empirical discrepancies affect user experience is not well understood. This leaves open the question of whether reducing these gaps would have a noticeable and desired impact on the speakers of these dialects.

Prior work, focused on the perspectives of African-American English speakers on Automatic Speech Recognition (Mengesha et al., 2021), has shown that directly asking subcommunities about their experiences with technology surfaces common problems and perceptions. Our work aims to extend this understanding to SAsE speakers. SAsE encompasses the varieties of English spoken in South Asia (Gargesh, 2019). While subvarieties (the largest being Indian English and Pakistani English) are defined by differing regions and local languages, they share common features (Kachru, 1986) and are often discussed in the context of Englishes in the South Asian Diaspora (Mahboob et al., 2008; Sharma, 2023; Masica, 2005). Our respondents are located in the United States and bilingual in English and at least one other South Asian language – corresponding most closely with sociolinguistic studies of Indian English speakers in California (Sharma, 2005a).

Our work conducts user-centric analyses of the use of language technology in SAsE, which has a large globally-spread speaker community (Gupta, 2008, 2010; Kachru, 1965) and previous empirical explorations in NLP research (Irvine et al., 2012; Sarkar et al., 2020; Demszky et al., 2021; Masis et al., 2022; Sun et al., 2023; Eisenstein et al., 2023; Ziems et al., 2023). We aim to understand both the degree of impact on SAsE speakers and how they interact with technology, adapt their speech, and express desired levels of support for their dialects. Furthermore, we assess which frustrations are unique to SAsE speakers by comparing to a control survey of SAmE speakers.

Our quantitative results indicate that SAsE speakers are significantly more likely to recall instances of language technology failures overall. Furthermore, responses about which specific technologies cause failures indicate that phonological variation,

[1]Benchmarks, Evaluation Code, and Full model predictions are released on Github and HuggingFace.

*Equal contribution, Listed in Alphabetical Order. Faye designed, administered, and analyzed survey responses. Will developed the intrinsic benchmarks and ran LLM evaluations. All authors decided on the project scope and to paper writing.

such as accents, may cause more universal issues for users, while written tech fails more uniquely for SAsE speakers. SAsE speakers indicate that they modify both their lexicon and syntax to make technology work better (e.g., avoid dialectal syntax, local terminology or "slang", and codeswitching), but that lexical issues are perceived as the most salient challenge.

We then develop two new benchmarks to assess the relevance of these challenges to state-of-the-art text-based NLP systems. As a *lexical* benchmark, we create a multiple choice evaluation of SAsE terms and their definitions, covering 317 loanwords and 724 stand-alone SAsE terms scraped from Wiktionary (Ylonen, 2022). As a *syntactic* benchmark, we create a minimal pair evaluation, similar to Warstadt et al. (2020); augmenting the 110 minimal pairs exhibiting grammar specific to Indian English introduced by Demszky et al. (2021) with synthetically generated negative examples of syntax that has not been attested in Indian English or Pakistani English (Kortmann et al., 2020; Ziems et al., 2023). As a control, we construct corresponding evaluations in SAmE for each benchmark.

In summary, we contribute the following:

1. **User-Centric Diagnostic Study of Failures:** We first assess the prevalence of failures for 78 SAsE and 97 SAmE speakers who met our criteria for analysis based on demographics, fluent languages, and responses to shibboleth terms from SAsE. In order to help researchers understand which aspects of SAsE are most salient to user interactions with technology, we further investigate user preferences and perceived challenges for 46 SAsE participants who opted-in to provide open-ended responses.
2. **Intrinsic Benchmarks of SAsE Knowledge:** We propose new intrinsic evaluations of the challenge categories identified by respondents. We assess understanding of 317 loanwords, 724 standalone dialect terms, and 110 syntactic features.
3. **Extensive Evaluations of LLMs:** We evaluate 8 families of open-source LLMs and 3 providers of closed-source LLMs for this understanding. We find that the disparities still exist across all categories of user frustration in the best-performing open-source models, while the most recently released GPT-4 model achieves near perfect performance.

2 Related Works

2.1 Attitudes Towards Dialectal English Use

English varieties spoken in "countries which were colonies of the English-speaking powers" (Platt, 1989) are often termed as "New" Englishes (Platt et al., 1984; Schneider, 2003), World Englishes (Kachru, 1992; Kachru and Nelson, 2006; Mesthrie and Bhatt, 2008), or Postcolonial Englishes (Schneider, 2007; Buschfeld and Kautzsch, 2017). In this categorization, SAsE is an umbrella term for some of the most widely spoken English varieties such as Indian English (Campbell and Grondona, 2008).

SAsE encompasses the varieties of English spoken in Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka (Gargesh, 2019). SAsE varieties have commonalities, but also a rich degree of variety due to differing local languages, cultures, and regions (Kachru, 1994); even within individual dialects such as Indian English there is variation (Kachru, 1965) from feature diffusion (Rickford, 1996; Nerbonne and Heeringa, 2010), other languages speakers are exposed to (Sharma, 2005a,b; Gargesh, 2019), and socioeconomic class boundaries (Sharma, 2023). Our study is limited to SAsE speakers who report also speaking an Indo-Aryan or Dravidian language, which limits our resulting sample to India, Pakistan, and Bangladesh.

Most related to our work is the study of perceptions and attitudes towards dialectal variation itself. Historically, many speakers of postcolonial dialects such as Indian English (Kachru, 1976; Hohenthal, 2003; Bernaisch and Koch, 2016), Singapore Colloquial English (Tan and Tan, 2008; Cavallaro and Chin, 2009), and Nigerian English (Olatoye, 2022) have indicated that grammar associated with SAmE or British English can indicate elevated social status, while grammar associated with local dialects can indicate solidarity. Notably, it may not be desirable for language technology to understand language used to indicate solidarity.

Some recent works have found that features of Indian English are increasingly preferred and viewed as more formal than SAmE features (Sahgal, 1991; Sharma, 2023), while others have still found that British English and SAmE are still viewed as indicators of elevated social status (Hohenthal, 2003; Bernaisch and Koch, 2016). Our work aims to understand how attitudes towards SAsE dialects are reflected in the context of language technology.

2.2 NLP Performance for English Dialects

Within NLP research, many works have explored discrepancies between SAmE and other dialects of English. African American Vernacular English (AAVE) is perhaps the most widely studied of these dialects. Performance drops for AAVE have been shown in POS tagging (Jørgensen et al., 2016b), sentiment analysis (Kiritchenko and Mohammad, 2018), dependency parsing (Blodgett et al., 2018), hate detection (Davidson et al., 2019), and seven NLU tasks defined by GLUE (Ziems et al., 2022b).

On a smaller scale, other works have demonstrated similar discrepancies for a broader range of global Englishes, including SAsE. Jurgens et al. (2017) demonstrate that language identification systems often misclassify global Englishes. In Faisal et al. (2021), the performance of ASR for question answering is shown to be poorer across eleven dialects of English (including Indian English) compared to SAmE. Finally, (Ziems et al., 2023) shows dialectal discrepancies for six dialects, including Indian English, for Machine Translation, Semantic Parsing, and Question Answering. Our benchmarks expand on these using intrinsic evaluations of both lexical and syntactic understanding of SAsE.

Mengesha et al. (2021)'s diary study of perspectives on ASR from 30 African-American English speakers highlights the distinction between purely empirical impacts and their practical impacts on the people who use systems. Distinguishing between the two creates a need to center work analyzing "bias" in NLP systems around the lived experiences of members of communities affected by these systems (Blodgett et al., 2020). Notably, it is not clear whether users *want* systems to understand their dialectal use. Our work aims to provide insights for developing dialectal NLP that meets users' desires, rather than needlessly risking reinforced harms on global English speakers through dual use (Kaffee et al., 2023; Held et al., 2023a).

3 Survey Design

To explore the wants of SAsE speakers within the realm of language technology, we extend prior user-centric surveying research (Mengesha et al., 2021) to understand user experiences with dialect usage and technology. Surveying directly addresses a gap in current NLP research, as we construct questions that allow us to analyze how empirical NLP failures with dialect data (Sarkar et al., 2020; Sun et al., 2023; Ziems et al., 2023) impact user perceptions and interactions. The survey results themselves serve as a foundation for constructing a robust failure framework, essential for guiding research to align with user needs and benchmarking models effectively against real-world user experiences.

Our survey aims to (1) quantitatively assess the differences in language technology failures between SAsE and SAmE speakers, and (2) gather qualitative feedback on user experiences and adaptations to better understand whether failure modes correspond to dialect usage. Before the study began, respondents were informed that "the purpose of this study is to understand how people use language to interact with technology." The survey starts with closed-ended questions to establish the occurrence and types of technology failures in English. This leads to open-ended questions exploring user perceptions and behavioral adaptations. By integrating both question types, we aim to minimize question ordering bias (Krosnick, 2018) and gain a comprehensive understanding of the aspects of a dialect that users employ or avoid when engaging with technology. Participants have the option to withdraw from the survey at any stage. We present the full survey in Appendix C.

3.1 Survey Sampling Procedure

After an initial pilot study (Appendix A), Prolific was used to run this survey due to its large and diverse participant pool, high data quality (Eyal et al., 2021; Douglas et al., 2023), balanced recruitment, and screening capabilities. This enabled us to filter for likely SAsE speakers based on bilingualism with English and fluency in least one other language common in South Asia. We were also able to filter for likely SAmE speakers by pre-screening for US-born participants who only speak English.

The survey was completed by 110 likely SAsE speakers and 150 likely SAmE speakers. We cross-check the validity of the pre-screening using both self-reported dialect information and shibboleth (Prokić et al., 2012) terms which distinguish SAsE and SAmE (*eggplant/brinjol, lentils/daal, elevator/lift*) (Appendix C). Likely SAmE speakers who self-identified as speaking dialects other than SAmE or responded with a SAsE aligned answer to any of the shibboleths were excluded. Likely SAsE speakers who passed the shibboleth check were included.

Surveys were hosted on the Qualtrics Platform and participants on Prolific were paid $15 per hour and the survey median time was 10 minutes. Sur-

veys were approved by the Institutional Review Board (IRB) at our institution and all researchers completed human-subjects research training.

3.2 Participant Profile and Representativeness

On Prolific[2], out of 302 eligible participants based on pre-screening, we recruited 113, and 78 passed our dialect verification checks. We also recruited 150 US-born participants who report only speaking English, 97 of whom self-identified as SAmE speakers and failed all shibboleth tests for SAsE as a control group. From the balanced sample in Prolific, our participants were US nationals self-reported as being from the US, India, Pakistan, and Bangladesh. Participants had a gender split of 47% women and 49% men (4% opted out of reporting gender) with a median age of 30. These demographics indicate that our survey has a slight skew to older men compared to general population statistics in these regions (CIA, 2017, 2021). Furthermore, as noted in the introduction, our results represent a sub-population with high contact with US English.

4 Survey Results

We analyze survey results by contrasting SAsE and SAmE speakers' quantitative responses and exploring the wants and frustrations of SAsE speakers through qualitative responses (shown here with the notation 'PX', shorthand for 'Participant X').

4.1 Prevalence of Misunderstandings

Our survey results (see Figure 1) show that a majority of both SAsE (75%) and SAmE (63%) participants recall instances when technology does not understand them well. Respondents were asked to mark or enter specific technologies they recalled experiencing issues with. These responses were coded as primarily speech-based (such as Voice Assistants or Automated Customer Service) or primarily text-based (such as Chatbots or Search Engines). SAsE speakers are significantly (+19%, P=0.026) more likely than their SAmE counterparts to list at least one written technology like ChatGPT, search engines, and Grammarly and significantly (-19%, P=0.012) less likely to list at least one spoken technology such as Siri, Alexa, and automated phone services. This finding indicates that the empirical disparities noted in prior works

[2]For a full demographic breakdown of our Prolific participants, see Appendix B, Table 3.

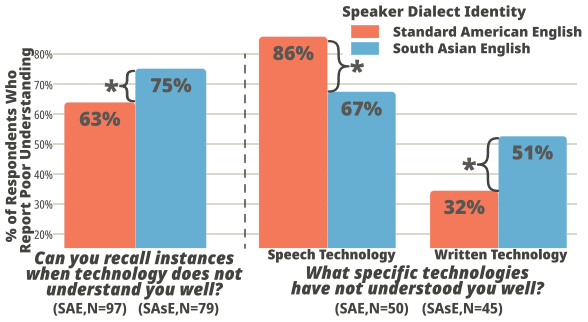

Figure 1: Survey responses to the questions *"Can you recall instances when technology does not understand you well?"* and *"What specific technologies have not understood you well?"*. * denotes significance at P<0.05 using a Barnard Exact test (Barnard, 1947).

on text-based NLP (Sarkar et al., 2020; Sun et al., 2023; Ziems et al., 2023) **create notably different user experience of language technology across dialect identity groups**.

However, this result does not indicate that written technology presents a larger challenge to SAsE speakers, as both groups more frequently (+54% SAmE, +16% for SAsE) list speech technology as a source of misunderstandings. It is unlikely that this response indicates that speech technology is *worse* for SAmE speakers than it is for SAsE speakers given prior empirical results (Javed et al., 2023). Instead, we argue these results indicate that issues with written technology are simply more salient for SAsE speakers. This could lead SAsE respondents to more frequently list only written failures when prompted, while issues from variation (e.g. accents) affect both SAmE and SAsE.

4.2 Perceived Causes of Failures

We further break down our survey analysis to uncover the main challenges SAsE speakers face when it comes to technology failures. We find three common challenges: (1) perception of technology failures with **stand-alone dialect words**, (2) perception of technology failures when **switching between languages**, and (3) perception of technology failures with **dialect features**.

While these identified challenges are not surprising, we analyze the frequency with which users cite each challenge in their responses and find that the challenge most frequently cited by users (failures with stand-alone dialect words) diverges from the challenges emphasized in existing research (i.e. syntactic failures (Ziems et al., 2023), switching between languages (Khanuja et al., 2020)). This

Challenge	Example Keywords	Occurrence
#1 Failures with stand-alone words	phrases, jargon, terminology, expressions, formal word, slang, yo, trend, different word, wrong word	43%
#2 Failures when switching between languages	foreign, other language, local language, bilingual, translate, punjabi, gujarati, urdu, hindi	18%
#3 Failures with colloquial dialect features	usage, formal language, dialect, diction, proper, standard, dialogue, colloquial	20%

Table 1: Reported challenges, corresponding keywords, and percentage of occurrences among users who responded to the open-ended questions, categorized by each challenge and its associated keywords.

points to a gap in current NLU research in addressing the wants of dialect speakers.

To analyze the frequency of these challenges in user responses, we examined the associated keywords, as shown in Table 1. By counting the occurrences of words in the survey responses, we identified patterns and grouped the most frequently appearing words. These keywords were then matched with their respective challenges based on the corresponding short-answer responses.

We also identified a common theme among participants linking technology failures and wanting technology to accommodate dialects:

> If you have a dialect that is not easy to understand, it will be harder to be understood by the tech you use. - P10
> I think technologies should be designed in a way that they are able to understand ever[y] dialect. - P18

4.2.1 Challenge 1: Stand-alone Dialect Words

Many participants (43%) report technology failures when using specific words from a given dialect, such as *buggy*, *ain't*, and *flat*. In response to these perceived failures, participants report:

> [I avoid using] some slang words. 'Buggy' instead of 'shopping cart' for example. - P2

Many participants make reference to the term *slang* and express that they avoid using slang when interacting with technology. Slang is linguistically defined as a speech variety (Zhou and Fan, 2013); it is used as "an instrument for in-group distinction," often intertwined with dialect usage, and a marker of colloquial speech (Drake, 1980). In specific registers, the use of register-level lexical variations (slang) can become completely intertwined with dialect (Drake, 1980; Chapel, 1998). Notably, participants indicate that they would *prefer* not to avoid slang and to interact with technology using more colloquial language: "I wish technology could understand the human phrases." - P22.

Participants' preference for colloquial language indicates that avoiding specific words is inconvenient, but perceived as leading to technology failures. Our keyword analysis in Table 1 highlights the major challenges users face when using stand-alone dialect words. This finding is notable as prior work primarily focuses on syntactic (Demszky et al., 2021; Sun et al., 2023; Ziems et al., 2023) or phonological variation (Faisal et al., 2021).

4.2.2 Challenge 2: Codeswitching

Participants also report general difficulties with switching between languages and the desire for technology to better adapt to these switches:

> I want to be able to speak bilingually with technology. - P7
> [I set Google Assistant] to the Gujarati language. I would sometimes ask for the weather [in English] and it would not understand me. - P6

The mixing of two languages in speech is a common feature of language use by bilingual individuals (Doğruöz et al., 2021), especially for Indian English speakers (Sharma et al., 2017; Rudra et al., 2019). The reported desire to switch between languages indicates an additional area where language technology may be failing to meet user needs. While codeswitching has been explored in the development of LLMs (Li et al., 2021), it also challenges language ID systems (Jurgens et al., 2017) which can lead to it being removed from pretraining data (Lucy et al., 2024). Multilingual models are not a catch all solution to codeswitching (Zhang et al., 2017), especially for SAsE where monolingual text often occurs in non-Latin scripts. In such cases, transliteration to the original script may be necessary to make use of representations

learned from monolingual text (Roark et al., 2020; Madhani et al., 2023; Held et al., 2023b).

4.2.3 Challenge 3: Register & Syntax

The final general pattern that emerged from survey responses was the tendency to avoid dialect patterns and features as a whole. Participants often cited that technology performs better when using *formal* over *informal* English:

> Language in for technology is so much more formal than spoken. - P19

Non-SAmE dialects are often described as informal (Hovy and Spruit, 2016), indicating participants may be avoiding dialect usage based on perceived failures. During the task-based portion of the survey (see Appendix C), 28% of participants dropped dialectal features when simulating an interaction with technology. For example, P33 changes "My childhood experience is still remembered by me." to "I still remember my childhood experience", dropping the feature of `object fronting` which has been attested in Indian English (Lange, 2012) and Pakistani English (Goetz, 2017) to make the technology work.

Users seem to be intrinsically aware of these discrepancies and are adapting their writing style accordingly, which suggests SAsE speakers are making extra efforts to overcome previously noted performance discrepancies resulting from syntactic variation (Ziems et al., 2022a, 2023). Instead of altering their writing to avoid perceived shortcomings, participants suggested technology should evolve to better suit user language:

> It should be technology that adapts to humans. - P24

5 Benchmarking LLMs on Challenges

While some survey respondents mention extremely recent services such as ChatGPT, most reference widely adopted technologies like customer service chat bots, search engines, and translation software. The connection between state-of-the-art research systems and those our respondents interact with on a regular basis is unclear. Therefore, we seek to empirically understand how the types of variation respondents report affects LLMs; a specific technology which is a recent focus for NLP research.

While benchmarks exist for Hindi-English codeswitching (Khanuja et al., 2020; Agarwal et al., 2023), syntactic differences (Ziems et al., 2023), and phonetic variation (Faisal et al., 2021), existing benchmarks do not cover all of the reported challenge categories and most notably omit standalone lexical variation, which is the largest issue mentioned by our respondents. To address this, we develop new intrinsic benchmarks tied to each of our challenge categories.

First, we create an intrinsic assessment of lexical understanding from Wiktionary (Meyer and Gurevych, 2012; Ylonen, 2022), covering 724 stand-alone terms representing Challenge #1 discussed in Section 4.2.1 and 317 loanwords from other South Asian languages representing Challenge #2 discussed in Section 4.2.2.

To assess Challenge #3 discussed in Section 4.2.3, we create a minimal pair syntactic language modeling evaluation in the style of Warstadt et al. (2020) with 110 sentences aligned between SAmE and Indian English (Demszky et al., 2021) augmented with aligned negative examples with syntax not attested in SAsE using rule-based transformations (Ziems et al., 2023).

We evaluate 8 series of open-source language models across all 3 assessments of reported challenges. For Challenges #1 and #2, we additionally evaluate on models from 3 industrial LLM providers but are unable to evaluate them on Challenge #3 due to limitations in their APIs. We show our results in Figures 2 and 3. Prompts used across all language models are provided in Appendix D.

5.1 Extracting SAsE Terms From Wiktionary

To evaluate lexical knowledge corresponding to Challenges #1 and #2, we gather terms from Wiktionary, a crowdsourced online dictionary. Wiktionary includes tags for lexical items that are affiliated with specific varieties of English, including seven variants of SAsE[3].

We use a Wiktextract (Ylonen, 2022), a machine-readable dump of Wiktionary, to gather all terms listed by users as Indian English (which encompasses 46 of 100 Pakistani English words and 9 of 30 Bangladeshi English words). To minimize inclusion of terms which may be irrelevant to speakers who use language technology today, we remove all terms categorized as archaic, obsolete, or historical by Wiktionary. This produces 1041 total nouns, verbs, and adjectives annotated as Indian English by Wiktionary contributors. We separate out 317 terms from substrate languages, such as

[3]List of Terms associated with SAsE on Wiktionary

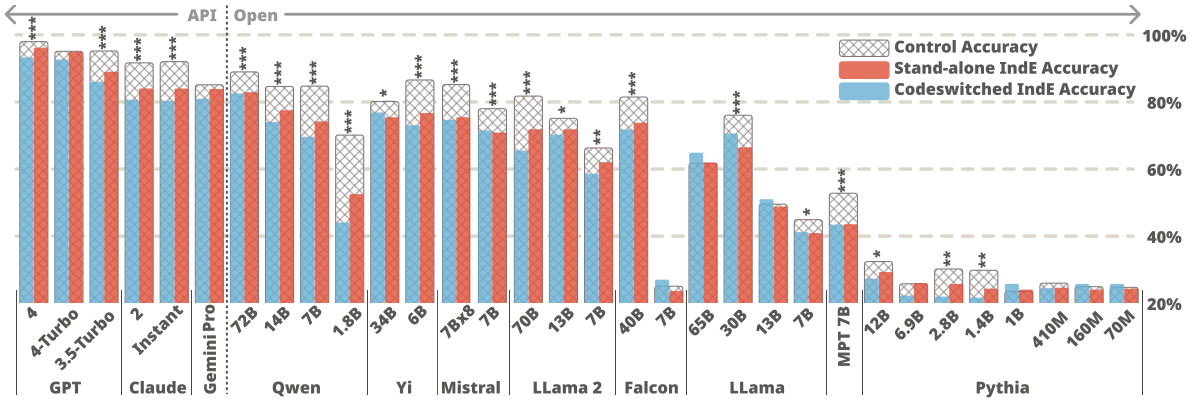

Figure 2: Results for Wiktionary Benchmarks of both SAsE and Unmarked Lexical Knowledge. *, **, and *** denote cases where overall performance is worse at P<0.05, P<0.01, and P<0.001 respectively by a Bootstrap test. Control accuracy is for terms without any regional affiliation on Wiktionary.

loanwords and calques, and assess Challenge #2 by intersecting this list with Wiktionary's list of English Borrowed Terms[4]. The remaining 724 terms, which are not marked as borrowed terms from another language, are used to assess Challenge #1. As a control point for comparison, we sample an equivalent set of 1041 terms that are not labeled with any particular regional dialect from the broader dataset.

We format these terms as multiple choice questions where the correct definition is placed alongside three incorrect definitions. The correct definition is the one provided by Wiktionary, while the incorrect definitions are randomly sampled from definitions of other terms. Each correct answer is assigned a different letter to prevent positional bias from over- or underestimating performance.

5.2 Evaluating Modeling of SAsE Syntax

While existing work has evaluated the functional effects of Indian English syntax on downstream tasks (Ziems et al., 2023), these assess the robustness of a model in the face of syntactic variation. We construct a more intrinsic benchmark of LLM understanding of acceptable lexical variation in Indian English, which is exhibited by respondents in their references to Challenge #3. Our evaluation follows the Benchmark of Linguistic Minimal Pairs (BLiMP) (Warstadt et al., 2020), comparing probabilities assigned to pairs of syntactically acceptable and unacceptable sentences with high lexical overlap. A language model with syntactic understanding should assign a higher probability to the acceptable sentence.

To develop a SAsE equivalent to BLiMP, we start with a dataset of minimal pairs between Indian English[5] aligned syntax and syntax aligned with SAmE or British English (Demszky et al., 2021). We then synthetically construct sentences that would be broadly unacceptable in both SAsE and in SAmE to serve as a negative baseline.

To do this, we first use eWAVE (Kortmann et al., 2020), a database of morphosyntactic features for varieties of English, to identify syntactic features whose absence has been attested by linguists in Indian and Pakistani English, confirming that experts in SAsE dialects would believe a sentence with this feature would be largely unacceptable. We then use a deterministic rule-based syntax transformation (Ziems et al., 2023) to convert each Standard American or British English example into an equivalent example which exhibits an unacceptable feature. We then sample a single unacceptable sentence for each example, providing a sentence with high lexical overlap but exhibiting a feature which has been verified by experts as unacceptable in both Pakistani and Indian English.

This gives us triplets of aligned sentences where one is produced according to syntax aligned with SAmE or British English, one is attested to occur in Indian English, and one is unacceptable in the SAsE covered by eWAVE. We use this to construct two exactly aligned minimal pair benchmarks, one where the correct sentences have Indian English syntactic features and one where they do not. In both cases, we use the same synthetically generated incorrect example as the negative.

[4]List of All English Borrowed Terms

[5]Some of the feature of Demszky et al. (2021) are not attested in Pakistani English and Bangladeshi English. However, overall Pakistani English and Indian English have a high degree of syntactic similarity with 43 out of 55 attested Pakistani English features attested in Indian English

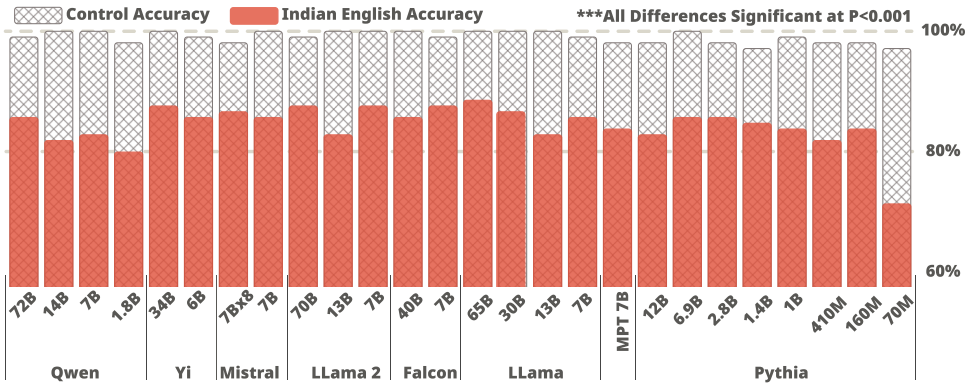

Figure 3: Results for Minimal Pair Benchmark of both Indian and SAmE Syntactic Knowledge. While even the smallest models consistently perform nearly perfectly on the SAmE control, even the largest models perform significantly (P<0.001) worse on the Indian English evaluation. Significance computed using a Bootstrap significance test.

In both setups, the expectation is that the model should assign higher probability to the sentence which demonstrates syntax which has been attested in Indian English than it does to the sentence which does not demonstrate any acceptable SAsE syntax. One shortcoming of this evaluation is that it relies on direct access language modeling probabilities, thereby limiting our evaluation to models where this is directly accessible.

5.3 Results

Challenges #1 and #2 results are shown side-by-side in Figure 2. Across open-access models, 14 out of 15 models which achieve greater than 60% accuracy on the control set perform significantly (P<0.05) worse on SAsE lexical knowledge overall. In general, models perform better on Challenge #1 with the exception of the first LLaMA models which perform better on loanwords (Challenge #2) at all scales. Promisingly, performance on Indian English terms is strongly correlated ($\rho = 0.98$) with performance on the control. This suggests that both issue categories are addressed to some degree by work thus far to improve models overall.

Furthermore, while 4 out of 6 industrial LLMs also have significantly (P<0.001) worse performance for SAsE, GPT-4 and GPT-4-Turbo both achieve over 90% accuracy on this benchmark. In a manual error-analysis, we find that 16 out of GPT-4's 57 remaining errors are terms that see limited usage online outside of historical documents. The next most common errors are definitions of slurs (5 errors), uncommon transliterations into Latin script (4 errors) and terms specific to agriculture (4 errors) and law (4 errors). The full error analysis is visualized in Appendix F.2.

Optimistically, this is a promising result indicating that the benchmark itself is achievable, but given the secretive nature of these models, it is unclear how to replicate such performance in open-access models. The prevalence of the significantly lower performance across evaluations of Challenges #1 and #2 provides quantitative support for surveyed user perceptions, even in recently developed systems.

Challenge #3 results are far more consistent across both model families and scales. Every model evaluated achieves near perfect results on the SAmE variant of the benchmark. Despite this, all models perform significantly (P<0.001) worse on our SAsE benchmark with the best performance being 89% accuracy achieved by LLama 65B.

Given that the same set of negative pairs is used for both control and SAsE evaluations, this drop is caused purely by the introduction of attested SAsE syntactic features. The consistency of this trend across scales of both model size and training data volume indicates that scaling is unlikely to provide intrinsic understanding of valid SAsE syntax.

Despite these exhibited performance drops, syntactic variation is much less frequently reported as a challenge by our respondents in Section 4.2.3. However, this may be unsurprising given evidence that syntactic understanding is frequently unnecessary to functional NLP (Pham et al., 2021). Still, for the long tail of cases which do require syntactic reasoning, these empirical results indicate that SAsE speakers may remain poorly supported (Papadimitriou et al., 2022).

6 Conclusions

Our work presents a user-centric diagnostic study of technology failures and further investigates whether these issues are pervasive in more recently developed LLMs. Our work studies the connection between the diminished empirical performance of NLP systems on inputs exhibiting SAsE linguistic features and the user experience of SAsE speakers. Concretely, we offer the following high-level takeaways: (1) While the majority of both groups recall issues with language technology, US-Based SAsE speakers do so 14% more often than SAmE speakers. (2) Differences in user experience go beyond accent. While spoken language technology more frequently causes issues for both groups, more SAsE speakers report issues with written language tech than their SAmE counterparts. (3) Users cite the most prescient pain-point as failures with standalone dialect words and report challenges with both words and syntax that have been attested in SAsE in free-form responses; users tend to remove such features to try and make technology work better. (4) Benchmark results support user perceptions, showing a performance dip in user identified challenge categories in recent LLMs. These results indicate that empirical differences in SAsE NLP performance create different perceptions of written language technologies for SAsE speakers. Therefore, language technologies must take linguistic variation into consideration, even for monolingual English systems.

Acknowledgments

We are grateful to Caleb Ziems, Camille Harris, Dora Zhao, Harshit Joshi, Jiaao Chen, Raj Sanjay Shah, and Rijul Magu for feedback, critique, and suggestions at various points in this work. We also thank Devyani Sharma for providing a valuable reading list at the start of this research. Computing resources for this project were in part provided through a Stanford Institute for Human-Centered Artificial Intelligence Google Cloud Credit Grant.

Limitations

Recruiting participants from Reddit is challenging due to the lack of demographic data available. Across Prolific and Reddit, while providing access to a diverse pool of participants, the study was constrained by the relatively small sample sizes available. Participants were mainly based in the United States. Our findings may not be generalizable to broader populations due to these constraints.

In regards to the study of SAsE specifically, both individual varieties and speakers are influenced by many different regional, economic, and linguistic backgrounds (Lange, 2012; Sharma, 2012). Our analysis surveys speakers who are bilingual in at least Hindi, Bangla, Urdu, Gujurati, Telugu, Tamil, and Malayalam, but further research may reveal differences in user preferences between variants of SAsE and within each variety itself. Further, we note that neither of the authors in this study speak a variety of SAsE. While we aimed to gather a diverse participant pool and research best methods for capturing SAsE user wants, this language limitation may have influenced our ability to fully understand and capture the perspectives of SAsE speaking participants.

Additionally, our work intentionally captures the perceptions of where technology is failing SAsE speakers in order to highlight the issues which are most valued by native speakers. However, in practice, NLP systems may be applied to users without their knowledge. Therefore, surveying about perceptions can easily undervalue the societal effects of pervasive, but less visible NLP systems which recommend content, target advertisements, and moderate platforms.

Ethics Statement

Our recruitment utilized the Prolific.Co platform. Notably, this meant that we did not recruit participants from outside of the United States for our collection of concrete issues. While our quantitative survey metrics capture a broader audience (excluding EU residents), this limits the perspectives which informed our data driven analysis of LLMs. As a human subjects survey, this project was reviewed and approved by the lead authors' Institutional Review Board.

References

2017. The world factbook central intelligence agency.

Anmol Agarwal, Jigar Gupta, Rahul Goel, Shyam Upadhyay, Pankaj Joshi, and Rengarajan Aravamudhan. 2023. CST5: Data augmentation for code-switched semantic parsing. In *Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants!*, pages 1–10, Prague, Czech Republic. Association for Computational Linguistics.

Ashley Amaya, Ruben Bach, Florian Keusch, and Frauke Kreuter. 2021. New data sources in social science research: Things to know before working with reddit data. *Social science computer review*, 39(5):943–960.

GA Barnard. 1947. Significance tests for 2× 2 tables. *Biometrika*, 34(1/2):123–138.

Tobias Bernaisch and Christopher Koch. 2016. Attitudes towards englishes in india. *World Englishes*, 35(1):118–132.

Terra Blevins, Robert Kwiatkowski, Jamie MacBeth, Kathleen McKeown, Desmond Patton, and Owen Rambow. 2016. Automatically processing tweets from gang-involved youth: Towards detecting loss and aggression. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2196–2206, Osaka, Japan. The COLING 2016 Organizing Committee.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

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, Johnny Wei, and Brendan O'Connor. 2018. Twitter Universal Dependency parsing for African-American and mainstream American English. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Melbourne, Australia. Association for Computational Linguistics.

Sarah Buschfeld and Alexander Kautzsch. 2017. Towards an integrated approach to postcolonial and non-postcolonial englishes. *World Englishes*, 36(1):104–126.

Lyle Campbell and Verónica Grondona. 2008. Ethnologue: Languages of the world. *Language*, 84(3):636–641.

Francesco Cavallaro and Ng Bee Chin. 2009. Between status and solidarity in singapore. *World Englishes*, 28(2):143–159.

Eble Chapel. 1998. Slang and sociability: In-group language among college students. by connie. *Journal of English Linguistics*, 26(3):247–265.

CIA. 2021. Median age - the world factbook.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

Dorottya Demszky, Devyani Sharma, Jonathan Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2021. Learning to recognize dialect features. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2315–2338, Online. Association for Computational Linguistics.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online. Association for Computational Linguistics.

Benjamin D Douglas, Patrick J Ewell, and Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between mturk, prolific, cloudresearch, qualtrics, and sona. *Plos one*, 18(3):e0279720.

Glendon Frank Drake. 1980. The social role of slang. In *Language*, pages 63–70. Elsevier.

Jacob Eisenstein, Vinodkumar Prabhakaran, Clara Rivera, Dorottya Demszky, and Devyani Sharma. 2023. MD3: The Multi-Dialect Dataset of Dialogues. In *Proc. INTERSPEECH 2023*, pages 4059–4063.

Peer Eyal, Rothschild David, Gordon Andrew, Evernden Zak, and Damer Ekaterina. 2021. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, pages 1–20.

Fahim Faisal, Sharlina Keshava, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2021. Sd-qa: Spoken dialectal question answering for the real world. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3296–3315.

Ravinder Gargesh. 2019. South asian englishes. *The handbook of world Englishes*, pages 105–134.

Sandra Goetz. 2017. Non-canonical syntax in south asian varieties of english: A corpus-based pilot study on fronting. *Zeitschrift für Anglistik und Amerikanistik*, 65(3):265–281.

Anthea Fraser Gupta. 2008. Singapore english. *The Handbook of World Englishes*, 7:351–368.

Anthea Fraser Gupta. 2010. Indian english. *The Handbook of World Englishes*, 7:203–222.

William Held, Camille Harris, Michael Best, and Diyi Yang. 2023a. A material lens on coloniality in nlp. *arXiv preprint arXiv:2311.08391*.

William Held, Christopher Hidey, Fei Liu, Eric Zhu, Rahul Goel, Diyi Yang, and Rushin Shah. 2023b. DAMP: Doubly aligned multilingual parser for task-oriented dialogue. In *Proceedings of the 61st Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada. Association for Computational Linguistics.

Annika Hohenthal. 2003. English in india: Loyalty and attitudes. *Language in India*, 3(5):1–107.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Ann Irvine, Jonathan Weese, and Chris Callison-Burch. 2012. Processing informal, romanized pakistani text messages. In *Proceedings of the Second Workshop on Language in Social Media*, pages 75–78.

Tahir Javed, Sakshi Joshi, Vignesh Nagarajan, Sai Sundaresan, Janki Nawale, Abhigyan Raman, Kaushal Bhogale, Pratyush Kumar, and Mitesh M. Khapra. 2023. Svarah: Evaluating english asr systems on indian accents.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2016a. Learning a POS tagger for AAVE-like language. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1115–1120, San Diego, California. Association for Computational Linguistics.

Anna Jørgensen, Dirk Hovy, Anders Søgaard, et al. 2016b. Learning a pos tagger for aave-like language. In *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Proceedings of the conference*. Association for Computational Linguistics.

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)*.

B Kachru. 1994. English in south asia. *The Cambridge*.

Braj B Kachru. 1965. The indianness in indian english. *Word*, 21(3):391–410.

Braj B Kachru. 1976. Models of english for the third world: White man's linguistic burden or language pragmatics? *Tesol Quarterly*, pages 221–239.

Braj B Kachru. 1986. The indianization of english. *English Today*, 2(2):31–33.

Braj B Kachru. 1992. World englishes: Approaches, issues and resources. *Language teaching*, 25(1):1–14.

Yamuna Kachru and Cecil L Nelson. 2006. *World Englishes in Asian Contexts*, volume 1. Hong Kong University Press.

Lucie-Aimée Kaffee, Arnav Arora, Zeerak Talat, and Isabelle Augenstein. 2023. Thorny roses: Investigating the dual use dilemma in natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13977–13998, Singapore. Association for Computational Linguistics.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

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

Jon A Krosnick. 2018. Questionnaire design. *The Palgrave handbook of survey research*, pages 439–455.

Claudia Lange. 2012. The syntax of spoken indian english. *The Syntax of Spoken Indian English*, pages 1–281.

Junjie Li, Jieyu Wu, and Richard Socher. 2021. Understanding code-switching in language models. *arXiv preprint arXiv:2109.04278*.

Li Lucy, Suchin Gururangan, Luca Soldaini, Emma Strubell, David Bamman, Lauren Klein, and Jesse Dodge. 2024. Aboutme: Using self-descriptions in webpages to document the effects of english pretraining data filters. *arXiv preprint arXiv:2401.06408*.

Yash Madhani, Sushane Parthan, Priyanka Bedekar, Gokul Nc, Ruchi Khapra, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Khapra. 2023. Aksharantar: Open Indic-language transliteration datasets and models for the next billion users. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 40–57, Singapore. Association for Computational Linguistics.

Ahmar Mahboob, Nadra Huma Ahmar, and Edgar W Schneider. 2008. Pakistani english: Phonology. *Varieties of English*, 4:244–258.

Colin P Masica. 2005. *Defining a linguistic area: South Asia*. Orient Blackswan.

Tessa Masis, Anissa Neal, Lisa Green, and Brendan O'Connor. 2022. Corpus-guided contrast sets for morphosyntactic feature detection in low-resource English varieties. In *Proceedings of the first workshop on NLP applications to field linguistics*, pages 11–25, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. 2021. "i don't think these devices are very culturally sensitive."—impact of automated speech recognition errors on african americans. *Frontiers in Artificial Intelligence*, 4:169.

Rajend Mesthrie and Rakesh M Bhatt. 2008. *World Englishes: The study of new linguistic varieties*. Cambridge University Press.

Christian M Meyer and Iryna Gurevych. 2012. *Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography*.

John Nerbonne and Wilbert Heeringa. 2010. Measuring dialect differences. *Language and Space: Theories and Methods. Berlin: Mouton De Gruyter*, pages 550–566.

Temitayo Olatoye. 2022. Attitudes of educated nigerians towards varieties of english. *Language Matters: Studies in the Languages of Southern Africa*, 53(1):81–102.

Isabel Papadimitriou, Richard Futrell, and Kyle Mahowald. 2022. When classifying grammatical role, BERT doesn't care about word order... except when it matters. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 636–643, Dublin, Ireland. Association for Computational Linguistics.

Thang Pham, Trung Bui, Long Mai, and Anh Nguyen. 2021. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1145–1160, Online. Association for Computational Linguistics.

John Platt. 1989. The nature of indigenized englishes: Interference — creativity — universals. *Language Sciences*, 11(4):395–407.

John Talbot Platt, Heidi Weber, and Mian Lian Ho. 1984. The new englishes.

Jelena Prokić, Çağrı Çöltekin, and John Nerbonne. 2012. Detecting shibboleths. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 72–80.

John R Rickford. 1996. Regional and social variation. *Sociolinguistics and language teaching*, pages 151–194.

Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J. Mielke, Cibu Johny, Isin Demirsahin, and Keith Hall. 2020. Processing South Asian languages written in the Latin script: the Dakshina dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2413–2423, Marseille, France. European Language Resources Association.

Koustav Rudra, Ashish Sharma, Kalika Bali, Monojit Choudhury, and Niloy Ganguly. 2019. Identifying and analyzing different aspects of english-hindi code-switching in twitter. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(3):1–28.

Anju Sahgal. 1991. Patterns of language use in a bilingual setting in india. *English around the world: Sociolinguistic perspectives*, pages 299–307.

Rupak Sarkar, Sayantan Mahinder, and Ashiqur KhudaBukhsh. 2020. The non-native speaker aspect: Indian English in social media. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 61–70, Online. Association for Computational Linguistics.

Edgar W Schneider. 2003. The dynamics of new englishes: From identity construction to dialect birth. *Language*, 79(2):233–281.

Edgar W Schneider. 2007. *Postcolonial English: Varieties around the world*. Cambridge University Press.

Alexander Shan, John Bauer, Riley Carlson, and Christopher Manning. 2023. Do "English" named entity recognizers work well on global englishes? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11778–11791, Singapore. Association for Computational Linguistics.

Devyani Sharma. 2005a. Dialect stabilization and speaker awareness in non-native varieties of english 1. *Journal of Sociolinguistics*, 9(2):194–224.

Devyani Sharma. 2005b. Language transfer and discourse universals in indian english article use. *Studies in Second Language Acquisition*, 27(4):535–566.

Devyani Sharma. 2012. Indian english. *The Mouton world atlas of variation in English*, pages 523–30.

Devyani Sharma. 2023. *From Deficit to Dialect: The Evolution of English in India and Singapore*. Oxford University Press.

Devyani Sharma, Alexander Bergs, and Laurel J Brinton. 2017. English in india. *The history of English*, 5:311–29.

Itamar Shatz. 2017. Fast, free, and targeted: Reddit as a source for recruiting participants online. *Social Science Computer Review*, 35(4):537–549.

Jiao Sun, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein, and Sebastian Gehrmann. 2023. Dialect-robust evaluation of generated text. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6010–6028, Toronto, Canada. Association for Computational Linguistics.

Peter KW Tan and Daniel KH Tan. 2008. Attitudes towards non-standard english in singapore 1. *World Englishes*, 27(3-4):465–479.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Tatu Ylonen. 2022. Wiktextract: Wiktionary as machine-readable structured data. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA).

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.

Yanchun Zhou and Yanhong Fan. 2013. A sociolinguistic study of american slang. *Theory & Practice in Language Studies*, 3(12).

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022a. 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, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022b. 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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

A Survey Pilot Study

A pilot study was performed on Reddit, to allow for targeted outreach potential (Shatz, 2017). We select participants from subreddits r/SampleSize and r/India based on prior usage in Reddit surveying (Amaya et al., 2021) and relevant population.

Based on feedback from the pilot, we revised the initial survey with a focus on gathering specific instances of adaptive behavior from SAsE speakers and more opt-in open-ended prompts for respondents to provide perspectives. Furthermore, due to the difficulty of confirming background information from Reddit we moved to Prolific for the final survey.

B Survey Demographics

	Gender			**Combined**
	Man	**Woman**	**Opt Out**	(N=78)
Total	49%	47%	4%	100%
Age (Verified)				
18-29	26%	54%	100%	42%
30-49	53%	38%	0%	43%
50+	18%	5%	0%	12%
Unknown	3%	3%	0%	3%
Median Age (in years)	34	28.5	23	30
Residence (Verified)				
US	100%	100%	100%	100%
Ethnicity (Self-Reported)				
Asian	87%	100%	100%	89%
White	10%	3%	0%	6%
Other	3%	9%	0%	5%
Country of Origin (Self-Reported)				
US	39%	35%	0%	36%
India	26%	30%	100%	31%
Bangladesh	18%	19%	0%	18%
Pakistan	11%	13%	0%	12%
Other (Taiwan, Saudi Arabia)	3%	3%	0%	2%

Table 2: Demographic Distribution of Prolific Survey Participants for the Sample of Speakers of SAsE.

	Fluent Languages (N=78)	**Primary Languages** (N=40)
Hindi	33%	20%
Bangla	26%	30%
Urdu	23%	20%
Spanish	12%	3%
Gujarati	9%	15%
Punjabi	8%	8%
Telugu	6%	8%
Chinese	4%	8%
Tamil	4%	0%
French	3%	0%
Other	3%	0%
Korean	1%	3%
Malayalam	1%	5%
Uzbek	1%	0%

Table 3: Distribution of Substrate Language Use and Familiarity reported by Prolific Survey Participants for the Sample of Speakers of SAsE.

C Survey Questions and Flow

Introduction:
By proceeding with this study, you attest that you are over 18 years of age. The purpose of this study is to understand how people use language to interact with technology. The questions are a mixture of multiple choice and short answer. This survey should take about 10 minutes. Your IP address is not recorded. Researchers do not have access to any personal information about you. Anonymous data from the study may be made public, but with NO link to you or your identity. The risks of this study are no greater than those involved in daily activities. You will not benefit from joining this study. We will comply with any applicable laws and regulations regarding confidentiality. To make sure that this research is being carried out in the proper way, the [Institution Name Redacted For Anonymity] IRB may review study records. The Office of Human Research Protections may also look at study records. Thank you for participating in this study; we appreciate your time and contribution to our research.
GDPR Compliance:
By proceeding with the following study, you attest that you are NOT a citizen or resident of the European Union (EU).

Dialect Self-report:
Displayed list of dialects. Which of the following best describe the dialect(s) of English you speak? (select all that apply)
Culture Checks:
Q1. *Displayed three stock photos of eggplants. Looking for participants to answer "brinjol", "brinjal", "bringol", "bringol", "aubergine", "begun", "bengan", "begun".*
Please write all the different names you use to refer to this vegetable.
Q2: *Displayed three stock photos of lentils. Looking for participants to answer "moth beans", "masoor", "massor", "daal", "dal", "chola", "masoor", "moshurdal".*
Please write all the different names you use to refer to this food.
Q3: *Displayed three stock photos of elevators. Looking for participants to answer "lift".*
Please write all the different names you use to refer to this object.

Recalling Pain Points:
Q1: Can you recall instances when technology does not understand you well? Yes / No
Skip To: Open-Ended Questions If Can you recall instances when technology does not understand you well? = No
Q2: When interacting with technology that does not understand you well, what language(s) have you used?
Q3: Are you ever able to make the technology work better? Yes / No / Sometimes
Display This Question: If Are you ever able to make the technology work better? ≠ No
Q4: What do you do to make the technology work better? (select all that apply)
Nothing / Reword/change your writing / Refresh the technology / Change the language / Other
Display This Question: If What do you do to make the technology work better? (select all that apply) = Reword/change your writing
Q5: Can you provide an example of how you change your writing to make technology work better?
Display This Question: If What do you do to make the technology work better? ≠ Reword/change your writing
Q6: Can you provide an example of your writing when interacting with technologies that did not understand you well?
Q7: Can you think of specific technologies that have not understood you well? Yes / No
Display These Questions: If Can you think of specific technologies that have not understood you well? = Yes
Q8: What specific technologies have not understood you well?
Q9: How would you categorize the technologies that have not understood you well? (select all that apply)
No Category/Not Applicable / Search engines / Social Media / Transportation Applications / Other
Task-Based Questions:
Q1: If you wanted to know how to cook eggs, how would you ask a friend? How would you ask a search engine (Google, Bing, DuckDuckGo, Yahoo, etc.)?
Q2: If you wanted to hear a song, how would you ask your friend to play the song? How would you ask a machine to play the song?
Q3: If you wanted directions to a restaurant, how would you ask a friend for directions? How would you ask a machine for directions?
Q4: If you wanted feedback on your writing, how would you ask a friend? How would you ask a machine?
Open-Ended Questions:
Q1: Generally speaking, how do you feel about technology's ability to understand you?
Q2: Would you prefer to be able to write differently than you do now when interacting with technology? How so? Can you provide a specific example?
Q3: Are there words or phrases you use when speaking that you avoid using when interacting with technology? If so, can you provide an example?
Q4: Are there any other insights or observations you have to share about your experience with technology and language understanding?

End of Survey:
Thank you for your participation! Your response has been recorded.

Figure 4: Survey Questions and Flow. Red text denotes survey skip logic. Blue text denotes participant answer options.

D Prompts

For both benchmarks, we use a single prompt across all models and for both the control and the SAsE versions of the results. Both prompts were written prior to running any evaluations, without further prompt engineering, and specify that the model should use knowledge of Indian English, since Indian English terms represent the majority of lexical items and all of the syntactic features.

For the lexical setup, we use the following multiple choice prompt, based on the best practices outlined in Ziems et al. (2022b):

```
Which of the following could \"{TERM}\" mean in
    Indian English when used as a {
    PART_OF_SPEECH}?
{OPTIONS A THROUGH D}
Answer:
```

For the syntactic setup, we compare the probabilities of the different sentences after the following prompt:

```
The following is an example of acceptable Indian
     English: "{SENTENCE}"
```

E Constructed Minimal Pairs

E.1 Challenge 1: Stand-alone Dialect Words

The elevator is stuck on the third floor.
The lift is stuck on the third floor.

At the grocery store I use a shopping-cart.
At the grocery store I use a buggy.

I want to go shopping.
I wanna go shopping.

What are some easy lentil recipes?
What are some easy daal recipes?

They are not going to the store.
They ain't going to the store.

Are you hungry right now?
Are yous hungry right now?

Do you want to drive?
Do you wanna drive?

Give me the salt please.
Gimme the salt please.

My apartment is being renovated.
My flat is being renovated.

E.2 Challenge 2: Codeswitching

How long should I cook an eggplant in the oven?
How long should I cook a brinjal in the oven?

I made over easy eggs for breakfast.
I made dim poach for breakfast.

Do you like fried eggplant?
Do you like begoon bhaja?

I have never tried lentils before.
I have never tried kichdi before.

E.3 Challenge 3: Register & Syntax

I need help with my writing, please give me feedback
I need help with my writing, please give me a feedback

How did you cook the eggs in the morning?
How did you cook egg in the morning?

I still remember my childhood experience.
My childhood experience is still remembered by me.

F Error Analysis

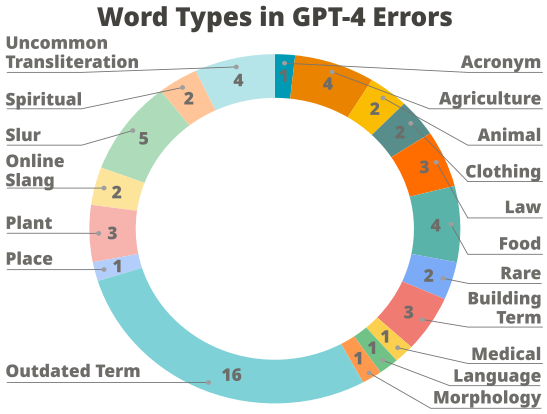

Figure F.2: Error Analysis of GPT-4