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., 2016; 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 raises 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 extends this line of work by surveying 78 South Asian English (SAsE) speakers and evaluating 11 families of open-access and industrial Large Language Models on new benchmarks to represent these tasks.

We contribute the following:

1. **User-Centric Dialect Study and Categorization of Main Challenges:** We investigate the preferences and perceived challenges faced by SAsE speakers when interacting with language technology, 78 of whom met our criteria for analysis (self-reported speaking a variety of SAsE and passed culture checks). We then code open-ended responses into challenge categories so that future research may focus on the pain points that are most salient to users.

2. **Intrinsic Benchmarks of SAsE Lexical and Syntactic Knowledge:** We propose new intrinsic evaluations of the challenge categories identified above. Our Lexical benchmark consists of 1041 terms, covering both loanwords and innovations, while our syntactic benchmark consists of 110 correct and incorrect minimal pairs. On these benchmarks, we find that disparities 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 Survey Design

Our survey aims to (1) quantitatively assess the differences in language technology failures between

[1]Benchmarks, Evaluation Code, and Full model predictions are released on Github.
*Equal contribution, Listed in Alphabetical Order.

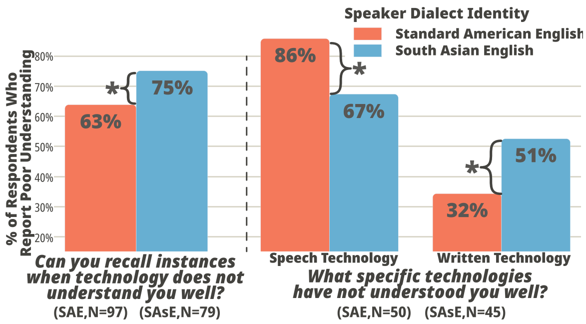

Figure 1: Survey responses to the listed survey questions. * denotes P<.05.

Challenge	Occurrence
#1 Failures with stand-alone words	47%
#2 Failures when switching between languages	14%
#3 Failures with dialect features	17%

Table 1: Reported challenges and percentage of occurrences in responses to open-ended questions.

SAsE and SAmE speakers, and (2) gather qualitative feedback on user experiences and adaptations to better understand whether failure modes correspond to dialect usage. We present the full survey in Appendix G.

3 Survey Results

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 also asked to mark specific technologies they recalled experiencing issues with. 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 on text-based NLP (Sarkar et al., 2020; Sun et al., 2023; Ziems et al., 2023) **create notably different user experiences of language technology across dialect identity groups**.

Perceived Causes of Failures

We further break down our survey analysis to the core challenges faced by SAsE speakers. 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**.[2]

4 Benchmarking LLMs on Challenges

While some respondents in our survey mention recent services such as ChatGPT, the connection between research representing state-of-the-art technology and those our respondents interact with in their day-to-day technology usage is unclear. Therefore, we construct benchmarks to assess the degree to which these challenges affect LLMs, a major recent focus area for NLP research.

Benchmark Construction

First, we mine a multiple-choice assessment of lexical knowledge from Wiktionary (Meyer and Gurevych, 2012; Ylonen, 2022), which includes community provided labels for terms primarily used in varieties of SAsE. We convert these terms into multiple choice questions for 724 stand-alone terms representing Challenge #1 and 317 loanwords representing Challenge #2 in Table 1.

We then create an evaluation of Challenge #3 discussed in Table 1 using linguistic minimal pairs (Warstadt et al., 2020) created by augmenting 110 sentences aligned between SAmE and Indian English (see Appendix C) (Demszky et al., 2021). The synthetically generated examples exhibit syntax not attested in SAsE using rule-based transformations (Ziems et al., 2023).

Evaluation Results

Across open-access models, 14 out of 15 models which achieve greater than 60% accuracy on the control set, perform significantly worse on SAsE lexical knowledge overall. 4 out of 6 industrial LLMs also have significantly worse performance for SAsE, but GPT-4 and GPT-4-Turbo both achieve over 90% accuracy.

Every model evaluated achieves near perfect results on the SAmE variant of the syntactic benchmark. Despite this, all models perform significantly worse on SAsE with the best performance being 89% accuracy achieved by LLama 65B.

Full results for these evaluations can be found in Appendix Figures E.2 and F.3.

5 Conclusions

These results suggest that even within English language technologies, dialectal variation plays a role

[2]The keywords used to code qualitative answers is included in Table A.2.

in the quality of service for different groups. Therefore, language technologies must take linguistic variation into consideration, even for monolingual English systems.

Limitations

Across Prolific and Reddit, the study was constrained by the relatively small sample sizes available. Additionally, both individual varieties of SAsE and speakers are influenced by different regional, economic, and linguistic backgrounds (Lange, 2012; Sharma, 2012). Further research may reveal differences in user preferences between variants of SAsE and within each variety itself. Further, we note that neither author speaks a variety of SAsE, potentially limiting our understanding of SAsE speaker perspectives.

Additionally, our work intentionally captures the perceptions of where technology is failing SAsE speakers to highlight issues which are most valued by native speakers. However, 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

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, 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.

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.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2016. 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.

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)*.

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

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.

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

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.

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

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.

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).

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 Keyword Details

Challenge	Example Keywords	Occurrence
#1 Failures with stand-alone words	phrases, jargon, terminology, expressions, formal word, slang, yo, trend, different word, wrong word	47%
#2 Failures when switching between languages	foreign, other language, local language, bilingual, translate, punjabi, gujarati, urdu, hindi	14%
#3 Failures with dialect features	usage, formal language, dialect, diction, proper, standard, dialogue, colloquial	17%

Table A.2: 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.

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 B.3: 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 B.4: Distribution of Substrate Language Use and Familiarity reported by Prolific Survey Participants for the Sample of Speakers of SAsE.

C Constructed Minimal Pairs

C.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.

C.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.

C.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.

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 Lexical Evaluation Results

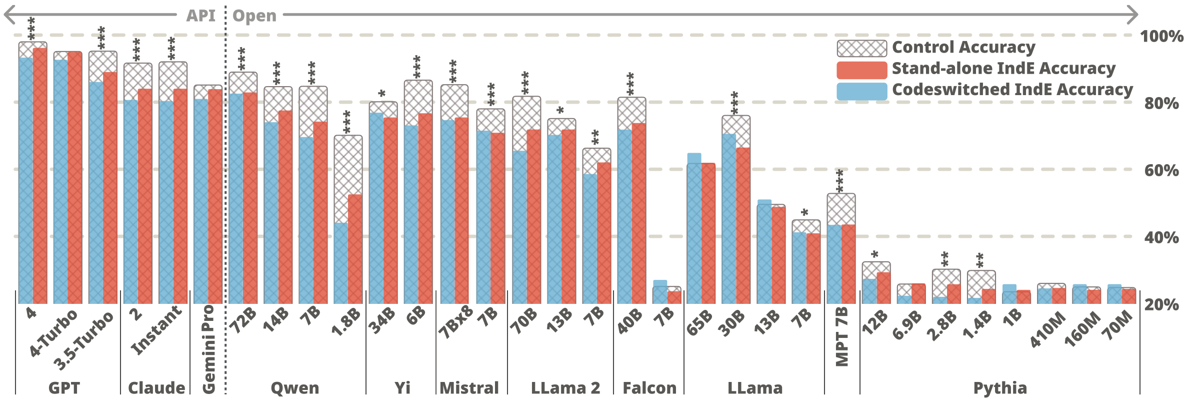

Figure E.2: Results for Wiktionary Benchmarks of both SAsE and Unmarked Lexical Knowledge. *, **, and *** denote cases where overall performance is worse at P<0.05, P<0.01, and P<0.001 respectively by a Bootstrap test. Control accuracy is for terms without any regional affiliation on Wiktionary.

F Syntactic Evaluation Results

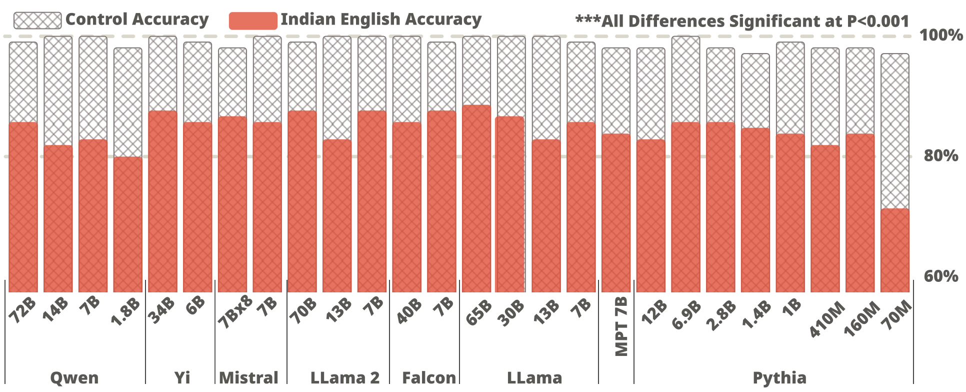

Figure F.3: Results for Minimal Pair Benchmark of both Indian and SAmE Syntactic Knowledge. While even the smallest models consistently perform nearly perfectly on the SAmE control, even the largest models perform significantly worse on the Indian English evaluation. Significance computed using a Bootstrap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

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