IMPROVING THE ROBUSTNESS OF NATURAL LANGUAGE PROCESSING TO DIALECTS AND LANGUAGE VARIATION

Thesis
Presented to
The Academic Faculty

By

William Held

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
Machine Learning Center
College of Computing

Georgia Institute of Technology

November 2025

IMPROVING THE ROBUSTNESS OF NATURAL LANGUAGE PROCESSING TO DIALECTS AND LANGUAGE VARIATION

Thesis committee:

Dr. Diyi Yang (Advisor)
Computer Science Department
Stanford University

Dr. Mark Riedl (Co-Advisor)
School of Interactive Computing
Georgia Institute of Technology

Dr. Zsolt Kira
School of Interactive Computing
Georgia Institute of Technology

Dr. Larry Heck
Schools of Electrical and Computer Engineering and Interactive Computing
Georgia Institute of Technology

Dr. Percy Liang
Department of Computer Science
Stanford University

ACKNOWLEDGMENTS

With the prevalence of AI and large language models (LLMs) in modern discourse, friends sometimes ask me whether I "knew" Natural Language Processing would be so impactful when I began doing research in the field. Universally, I tell them that I had no clue! I picked the field, in large part, because, even as an outsider, it seemed full of amazing, intellectually curious people. From my first course in NLP, taught by my undergraduate advisor Professor Nizar Habash, I have been unbelievably lucky to collaborate with many such researchers.

Firstly, my Ph.D. would not have been possible without my advisor, Professor Diyi Yang. Unsurprisingly given her research, Diyi has shaped my understanding of the human side of research deeply. While this surfaces in my work through questions of how we can shape human-centered issues into hard technical problems, the most important lesson Diyi has taught me is how to build a happy and healthy research community. I joined Diyi's Social and Language Technology (SALT) Lab in its third year, and it has been a joy to see it grow and flourish under her leadership. Diyi has gone above and beyond to make my Ph.D.—as well as those of all her other students—as intellectually free as could be asked for. Despite immense turmoil in our field and in the broader world, Diyi was always in my corner, making sure I had everything I needed to drive my research vision forward.

I had the unusual privilege of spending time at two universities during my Ph.D. I spent the first two and a half years at Georgia Tech, where I was able to build my confidence in machine learning theory and practice. I am fortunate to have two committee members on my thesis who played key roles in that process as teachers: Professors Zsolt Kira and Larry Heck. I was further lucky to have a generous community at Georgia Tech more broadly, exemplified by my co-advisor Professor Mark Riedl, who did the unglamorous but essential work necessary to make Diyi's transition to Stanford smooth for me logistically within Georgia Tech. Mark took this additional work on without as much as a second

thought, and I am deeply appreciative of the work this required of him. More recently, I have spent my time with the Stanford NLP Group as a visitor, where I have been lucky to work with Professor Percy Liang on both my own research and on envisioning a future where LLM research itself is more open and collaborative. In a field that is increasingly competitive, I am so grateful to have benefited from the investment of all these mentors, as well as many folks in both departments—including Judy Hoffman, Mike Best, Danfei Xu at Georgia Tech, and Dan Iter, David Hall, and Dan Jurafsky at Stanford—who have guided and helped me in various ways throughout my Ph.D.

Looking back on my research trajectory, my collaborations with industry researchers always gave me new ways to update my understanding of the core challenges in building useful tools for real users and influenced how I approached my work upon returning to the SALT Lab. I am especially appreciative of my mentors within industry: Chris Hidey at Google, Shruti Bhosale at Meta, and Todor Mihaylov at Meta.

Most of all, I was lucky throughout my Ph.D. to have a truly great group of collaborators in the SALT Lab. I had the privilege of co-authoring work with all the members of the SALT Lab who were at Georgia Tech when I joined—Jiaao Chen, Caleb Ziems, Camille Harris, Yanzhe Zhang, and Omar Shaikh—and I continue to learn from all of the folks who have joined us at Stanford—Ryan Louie, Hao Zhu, Weiyan Shi, Jing Huang, Yijia Shao, Chenglei Si, and John Yang. Finally, I likely learned the most—both about research and about myself—through mentoring fantastically talented undergraduate and master's students, including Michael Ryan, Yanchen Liu, Faye Holt, Woody Gan, Vyoma Raman, and others.

Without my partner Olivia, who is unimaginably patient with my research rants and ravings, I likely would have lost my mind or my health at some point during this process. Finally, and perhaps most importantly, it is only thanks to my family—my parents, Lani and Bruce, and my siblings, Sergei and Nikki—that I had the capabilities, resources, and belief in myself to contribute to such an amazing research area.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION AND BACKGROUND

1.1 Dialects, Language Variation, and Natural Language Processing

Variation is a fundamental aspect of human language, manifesting across geographical boundaries, social strata, time shifts, and communicative contexts. While the Natural Language Processing (NLP) community has historically focused on relatively homogeneous samples of language [1], the adoption of user-facing language technologies—from machine translation to LLM-enabled chatbots— has increased recognition of the importance of making models robust to language variation [2, 3, 4]. Beyond delivering better systems to users, language variation represents a significant class of real-world distribution shifts [5], an area of interest to the broader machine learning community.

Language variation emerges through localized shifts in individual utterances, originating from both practical necessities and unpredictable community preferences. Within a single conversation, variation is constrained by intelligibility, as each utterance must be understood by others to be useful. However, when linguistic communities experience prolonged separation, variations can develop beyond mutual intelligibility as groups establish internally coherent norms. Categories of linguistic variation are generally defined by the type of barrier which caused said separation.

Diachronic variation refers to changes in language over time, as successive generations introduce linguistic innovations and languages evolve through contact processes [6]. Diatopic variation encompasses differences across geographic regions, manifesting as dialects and regional varieties [7]. Diastratic variation reflects social stratification factors such as education level, socioeconomic status, and professional background [8, 9]. Diaphasic variation, also known as register variation, describes systematic linguistic differences

across communicative contexts, including levels of formality, medium, and communicative purposes [10]. These dimensions interact dynamically, shaping how individuals speak in context-dependent ways [11, 12].

As might be clear from the above, language variation has been studied in a wide variety of forms in NLP from entirely separate language families [13] to subtle differences based on the target audience of an utterance [14]. For both practical and theoretical reasons, my research has focused primarily on language variations which are broadly defined as World Englishes [15]. Practically, this research direction is motivated by the position of English as a global language with natively spoken varieties on every continent except Antarctica. Despite the elevated level of support within NLP for Standard American English (SAE), many English speaking communities still may avoid language technology "simply because local spoken varieties of English are still not well supported" [16]. From a theoretical perspective, many dialects of English are largely mutually intelligible with SAE, providing evidence that dialectal robustness is a feasibly addressable form of domain shift.

My research has focused on measuring the impacts of dialect bias in NLP for World Englishes and building methods to mitigate these biases. In the final phase of my thesis, I propose a course of work to develop statistical models of the interactions between Large Language Model training data, model size, and dialect disparities. My goal is to contextualize my work in an NLP landscape that has been transformed by the scaling of Pretrained Language Models (PLMs) toward general purpose technology.

1.2 Mixed-Methods Studies to Identify and Measure Dialect Disparities (Chapter 2)

In the first part of my thesis, I will cover my work which has aimed to cement the connection between the use of known features of English dialects and negative impacts on both empirical performance metrics and downstream user experience. While prior works on dialect disparities exist, my work has addressed key limitations in this area. Firstly, in evalu-

ations focused purely on naturally occurring data, the causes of dialect disparity were often unclear due to the existence of confounding variables such as topical differences across data or style confounds unrelated to dialect specifically such as formality. Second, prior works often focused solely on a single dialect — frequently African American Vernacular English (AAVE) — with less focus on the range of global English dialects. Finally, prior work had minimal focus on establishing the impacts of empirical performance disparities on the user experience of real users.

First, I will introduce the Multi-VALUE toolkit [17], a framework which allows for carefully controlled stress-tests of the impacts of language variation on NLP systems; extending the previous VALUE framework [18] designed for AAVE to 50 English dialects. Unlike work focused on naturally occurring data, Multi-VALUE enables the creation of synthetically generated dialectal counterfactuals, enabling results which establish causal links between specific linguistic features and performance differences.

Then, I will explore my work analyzing the impacts of dialect use on user experience for dialect speakers [19]. By performing a controlled study between SAE Speakers and English speakers who speak varieties of English originating in South Asia, I show that South Asian English speakers report significantly more misunderstandings with Language Technology, they modify their natural linguistic patterns to accommodate the limitations of NLP systems, and they express a desire to utilize their dialect to a greater degree when interfacing with language technologies. We then constructed targeted benchmarks for issues that users expressed, validating their existence across 11 families of LLMs.

1.3 Methods for Rapid Dialect Adaptation through Finetuning (Chapter 3)

Leveraging my evaluation methodologies, the second section of my thesis will focuses on my work developing methods to address the identified disparities. Specifically, I have focused on methods which rapidly adapt Pretrained Language Models (PLMs) to new dialects

of English in a task-agnostic fashion. Prior to my work, machine learning methods for dialectal English NLP were task-specific, relying on manually annotated dialect data [20, 21], weak-supervision [22, 23], or data augmentation [18, 17].

However, LLMs have become an increasingly general-purpose NLP tool during my PhD, creating a long-tail of use cases. Across the range of tasks, the cost of additional training is likely to prevent practitioners from adopting a new dialect mitigation, especially since many practitioners undervalue secondary metrics such as robustness [24]. Instead, my work has focused on *task-agnostic* adaptations which use alignment losses to optimize demographic parity [25] at the representation level of models.

I first developed the core optimization method for this form of adaptation while working on task-oriented parsing for Hindi-English and Spanish-English code-switching[26]. Later, I combined the optimization principles with computationally efficient finetuning techniques to create a plug-and-play framework called Task Agnostic Dialect Adapters (TADA) [27]. Finally, I advised two extensions to TADA which further simplified robustness improvements by enabling zero-shot transfer to *unseen* dialects by developing novel machine learning methods to incorporate structural features of dialects into the neural network architectures used for model adaptation [28, 29].

1.4 Forecasting and Improving Dialect Robustness During Pretraining (Chapter 4)

As Large Language Models (LLMs) have scaled, their aggregate performance has followed predictable *scaling laws*—empirical power-law relationships between compute, data, model size, and loss [30, 31]. These relationships have been validated for overall accuracy and cross-task generalization, but their implications for *linguistic subpopulations* remain largely unknown. A central open question for fairness and sociolinguistic NLP is whether scaling alone can eliminate disparities across dialects or whether inequities persist despite global performance gains.

To address this, I conduct the first systematic empirical study of *dialectal scaling dynamics* in pretrained language models. Using a compute-controlled series of models trained under consistent optimization regimes, I measure how performance across English dialects changes as a function of model size and data volume. The analysis integrates both classical scaling law estimation and *relative scaling laws*, which characterize how subgroup-to-baseline loss ratios evolve with scale.

Results show that while overall performance improves monotonically with scale, disparities relative to Standard American English (SAE) do not vanish. Dialects that are better represented in pretraining data—such as Canadian or Singapore English—exhibit neutral or slightly positive scaling slopes, implying proportional improvement with scale. In contrast, underrepresented varieties—such as Sri Lankan and Nigerian English—display shallower or negative scaling slopes, meaning that performance gaps *widen* as models grow. Crucially, these trends seem to arise primarily from *capacity scaling* (increasing parameters) rather than *data scaling* (adding tokens): holding data constant while varying model size changes relative performance, whereas increasing data alone does not. Across International Corpus of English (ICE) dialects, the correlation between estimated scaling slopes and English-speaking internet population size is strong (Pearson $R{=}0.82$–0.84, $p{<}0.005$), suggesting that online representation during pretraining modulates scaling efficiency.

These findings demonstrate that larger models are not "universal equalizers." Scale raises mean accuracy but leaves subgroup disparities largely intact or even magnified. Achieving dialect robustness therefore requires intentional design—through balanced pretraining mixtures, synthetic augmentation (e.g., Multi-VALUE [17]), or task-agnostic adaptation methods (e.g., TADA [27])—rather than relying on scale alone. By situating dialect disparities within the quantitative framework of scaling laws, this chapter provides a predictive model for how linguistic inequities evolve with pretraining scale and establishes that fairness in language technology is an active variable, not an automatic byproduct of bigger models.

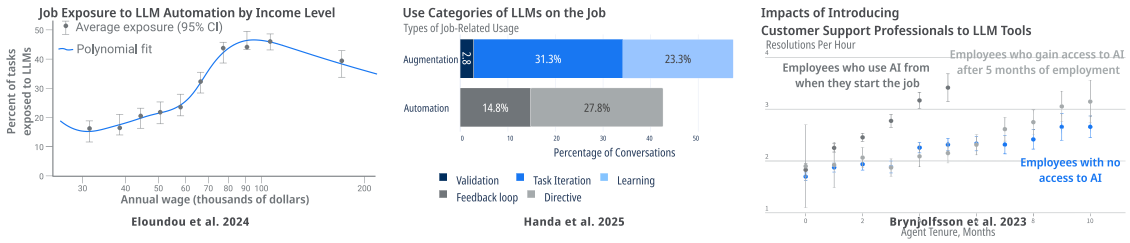

Figure 1.1: **Broader Research on LLMs as General Purpose Technology.** (a) Language models could theoretically impact a huge share of the economy, including many tasks traditionally viewed as high income[32]. (b) Current usage of LLMs, and NLP more broadly, is largely augmentative to human capabilities[33]. (c) When adopting LLMs in workplace use, there are early signs that LLMs can make employees more productive[34].

1.5 Position within the Broader NLP Development

The importance of dialectal robustness in NLP has, to me, never been more clear than it is today. As shown in Fig. 1.1, there is an increasing body of work which shows that LLMs are being adopted broadly in the workplace. While this research is early, much of this work seems to point to the most likely outcome being that LLMs can be used, much like most previous technology[35], to improve the productivity of existing workers. This is especially important given that a very broad section of the economy is exposed to this potential use of LLMs [32], and especially relatively high-earning jobs.

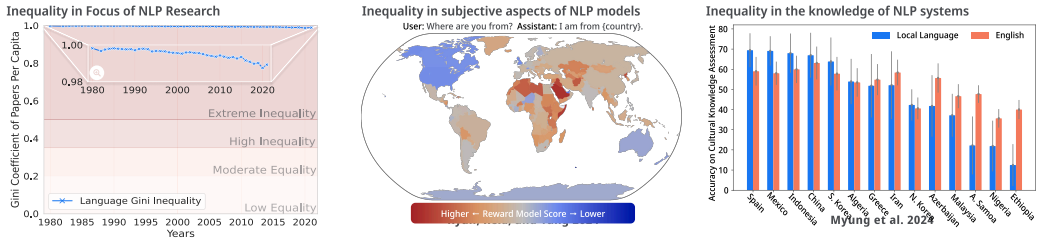

Figure 1.2: **Broader Research on LLMs as Cultural Technology.** (a) NLP research is distributed extremely unequally across languages and cultures[36]. (b) LLMs exhibit clear biases in opinion-based questions about the world[37]. (c) LLMs exhibit less knowledge about different regions, both in English and in local languages[38].

However, unlike many previous general-purpose technologies, LLMs and NLP systems more broadly are also *cultural* technologies whose performance can be meaningfully impacted by the identity of the user of a system. As shown in Fig 1.2, my own and others work

has shown that this surfaces in many aspects of both NLP investment and performance. The combination of these trends, where tools increasingly general-purpose but unequally performant for different groups of people has the potential to increase inequality rather than decrease it. It is this potential which motivates my area of work.

1.6 Thesis Statement

English as a global language—spoken by billions across continents—is rich with systematic variation. Yet most language technologies are trained on and optimized for Standard American English, leading to consistent performance gaps and usability barriers for other dialect communities. In this work, I establish empirical evidence for these disparities through controlled experiments and user studies spanning multiple English varieties. I show that these gaps persist even as models scale, demonstrating that increased capacity alone does not produce dialect robustness. I then develop computationally efficient adaptation methods that close these gaps without requiring dialect-specific annotations for each task. Collectively, my findings show that dialect disparities in NLP are measurable, persistent under scale, and resolvable through targeted robustness and fairness methods.

CHAPTER 2

MIXED-METHODS STUDIES TO IDENTIFY AND MEASURE DISPARITIES

2.1 A Toolkit for Stress-Testing Dialect Disparities

Even before my work, there was a significant amount of research studying dialect disparity in NLP for AAVE across many tasks. Performance gaps had been documented for language use from Black American users and on social media in predominantly Black regions of the United States across hate speech classification [39, 40, 41, 42, 43], NLI [18], dependency parsing, POS tagging [20, 22], and downstream applications [44]. However, there did not exist a systematic exploration of robustness across multiple Englishes.

In my co-first author work with Caleb Ziems, Multi-VALUE [17], we expanded the VernAcular Language Understanding Evaluation (VALUE) framework of Ziems *et al.* [18] to allow for analysis of dialect robustness in 50 English dialects through the use of 189 controllable perturbations. Multi-VALUE offered the following advantages

1. **Interpretable:** enables causal analysis through controllable counterfactuals.
2. **Flexible:** designed to model new and evolving dialects by adjusting dialect density and makeup, which is leveraged in later methods work (See 3.2.3).
3. **Scalable:** allows users to analyze new tasks without additional human annotations.
4. **Responsible:** vetted by native speakers to ensure gold standards and synthetic data are dependable for ongoing research.
5. **Generalizable:** moves the field beyond single-dialect evaluation, which allows researchers to draw more transferrable findings about cross-dialectal NLP performance.

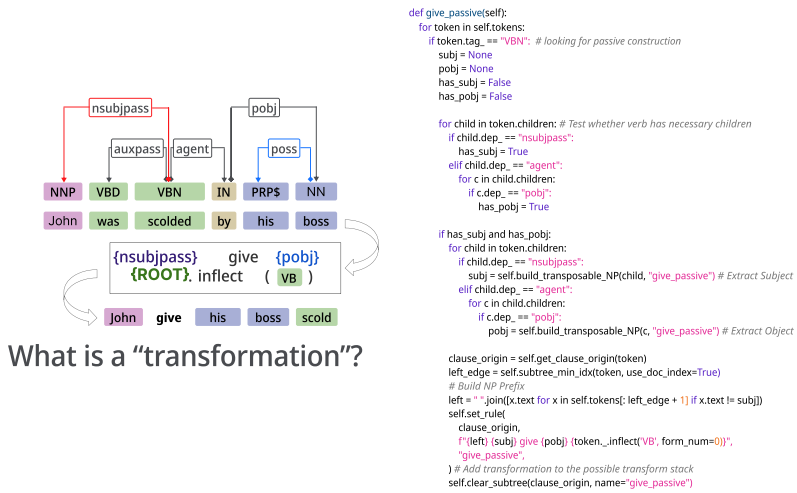

Figure 2.1: **Example of Linguistic Transformation** in practice. Held *et al.* [17] instantiates high-level linguistic patterns as low-level operations over syntactic parse trees.

2.1.1 Transformation Construction

Multi-VALUE is a practical instantiation of the linguistic knowledge accumulated in the Electronic World Atlas of Varieties of English [45] — a linguistic repository documenting syntactic variation in global Englishes accumulated by experts in each individual English variety. While eWAVE describes syntactic features in free-text, Multi-VALUE implements these features by analyzing inputs with POS tagging, inflectional analysis, and dependency parsing via `spaCy` [46] and `stanza` [47] libraries and then manipulating the results with deterministic logic. Following the eWAVE organizational scheme, Multi-VALUE constructs perturbations across 12 grammatical categories: (1) Pronouns, (2) Noun Phrases, (3) Tense and Aspect, (4) Mood, (5) Verb Morphology, (6) Negation, (7) Agreement, (8) Relativization, (9) Complementation, (10) Adverbial Subordination, (11) Adverbs and Prepositions, and (12) Discourse and Word Order.

With implementations for features that have been documented across all English varieties, an individual dialect is defined as a feature vector over all possible features. For each feature, eWAVE defines whether the feature is pervasive, neither pervasive nor rare, rare, and or absent in a particular dialect. Multi-VALUE treats these categories as probabilities

Feat.	Acc.	Feat.	Acc.	Feat.	Acc.	Feat.	Acc.
10	97.4	67	99.1	128	92.7	173	87.5
39	99.7	70	92.9	130	92.9	175	83.3
40	99.8	71	98.8	132	87.7	193	88.7
42	98.1	88	99.4	133	99.5	216	99.7
43	93.2	96	95.5	154	92.9	220	99.4
49	99.6	99	94.7	155	81.8	221	86.7
56	97.3	100	99.9	165	99.1	224	99.5
60	99.6	121	91.8	170	94.9	227	91.2
63	99.0	126	92.3	172	90.0	228	99.8

Feats.	Acc.
3, 9, 11, 14, 15, 16, 26, 29, 33, 34, 41, 45, 47, 55, 57, 58, 59, 61, 62, 64, 66, 77, 78, 79, 80, 81, 86, 101, 106, 117, 119, 123, 131, 134, 145, 146, 149, 159, 174, 179, 191, 194, 198, 203, 204, 205, 206, 207, 208, 209, 214, 223, 226, 232, 235	100.0

Table 2.1: **Accuracy of 92 perturbation rules** according to majority vote with at least 5 unique sentence instances. Seventy four rules have >95% accuracy, while sixteen have accuracy in [85,95), and only two are <85% accurate, demonstrating the reliability of our approach.

Model		Test Dialect		
Base	Train Set	SAE	ChcE	IndE
BERT	SAE	77.2	76.7 (-0.5%)	72.3 (-6.7%)$^{-}$
	Multi	76.2 (-1.2%)	76.1 (-1.4%)	75.0 (-2.9%)$^{+-}$
	In-Dialect	77.2	76.5 (-0.9%)	75.1 (-2.7%)$^{+-}$
RoBERTa	SAE	81.8	81.6 (-0.2%)	77.7 (-5.2%)$^{-}$
	Multi	80.6 (-1.5%)$^{-}$	80.5 (-1.6%)$^{-}$	79.7 (-2.7%)$^{+-}$
	In-Dialect	81.8	81.6 (-0.2%)	80.5 (-1.6%)$^{+-}$

Table 2.2: **CoQA Evaluation:** F1 Metric on each gold development set of the CoQA benchmark. $^{-}$ and $^{+}$ respectively indicate significantly ($P < 0.05$) worse performance than SAE↦SAE and better performance than SAE↦Dialect by a paired bootstrap test.

of the feature occurrence with 100% probability for pervasive features; 60% for neutral features; 30% for rare features; and absent features being skipped. Finally, to transform a particular input to be aligned with a particular dialect transformations are applied sequentially. Multi-VALUE covers 189 of 235 features documented in eWave, with no dialect having less than 80% of its features implemented. The remaining unimplemented features require information which is not accessible from morphosyntactic parsing, such as mood, aspect, or conversational context such as group size.

2.1.2 Transformation Validation

Since Multi-VALUE is a system of synthetic transformations, a key aspect of the work is validating that the system generates text which is plausible and grammatical to native speakers of a particular dialect. This is a key differentiating feature from work using unvalidated synthetic features without clear correspondence to real world variation [48].

To verify the reliability, we recruited English speakers on Amazon Mechanical Turk. We first asked them to self-report their spoken dialects and then and administered a survey about their grammaticality judgements for manually constructed sentences demonstrating

attested features from eWAVE. If their grammaticality judgements align with their self-reported dialects, they are added into the annotator pool.

Using this process, we recruited 72 annotators across 10 English dialects who then labeled the accuracy of individual perturbations corresponding to features which exist in their native dialects. Perturbation accuracies are given in Table 2.1. Since 55 rules have 100% accuracy, with all rules maintaining accuracy above 81%, Multi-VALUE is a well-validated synthetic variation testing environment.

2.1.3 Initial Multi-VALUE Analyses

While Multi-VALUE can apply to any task with free-form text, our work focused on evaluating three tasks in particular: conversational question answering, semantic parsing, and machine translation. All three are user-facing tasks where language variation may hinder users' access to information, resources, and/or the global economy [49, 50].

For brevity in this proposal, I focus on our results for conversation question answering based on CoQA from Reddy *et al.* [51]. We study this task because the conversational nature of the questions, which include references to previous content, allows dialectal errors to compound. To transform the publicly available training and development sets, we perturb only questions, simulating the setting where the user submits queries in a low-resource dialect while the system is expected to respond in SAE. For this task, we further cleaned the Multi-VALUE constructed test data by allowing human annotators to edit system outputs for both Chicano English (CHcE) and Indian English (IndE) to improve naturalness.

We show the results on this gold standard data in Table 2.2 for the BERT[52] and RoBERTa[53] Base models. Chicano English, which is similar to SAE, does not have significantly worse performance. However, for Indian English, models have significantly worse results (-6.7% BERT, -5.2% RoBERTa). We then leverage Multi-VALUE as a augmentation tool and train on a synthetic pseudo-dialect using random permutations of all feature options available. This synthetic data augmentation significantly improves results

on real Indian English (+3.8% BERT, +2.5% RoBERTa) data.

In the complete work, we performed similar evaluations on both generative models, such as T5[54] and BART[55], for semantic parsing on the SPIDER benchmark [56] and models specified for machine translation [57] on the WMT-19 benchmark [58]. In all of these settings, we found similar patterns — namely that the more distant a dialect was syntactically from Standard American English the more that model performance degraded on these dialects.[1]

2.2 Surveying Dialect Speakers

While Multi-VALUE had clear empirical findings, these do not, a priori, confirm our motivations in exploring dialect are sound since empirical differences may not surface as user experience impacts. As such, in my subsequent research I was interested in getting insight into dialectal NLP from the perspective of users.

Prior work, focused on the perspectives of African-American English speakers on Automatic Speech Recognition [59], had shown that directly asking subcommunities about their experiences with technology is a simple but effective way to surface problems and perceptions. Building on Multi-VALUEs finding of significant differences in Indian English, myself along with my co-first author, undergraduate Faye Holt, decided to perform a user survey to better understand the perspectives of speakers of South Asian Englishes (SAsE), the family of English varieties spoken in South Asia [60].

Despite South Asia having an enormous English speaking community [61, 62] and extensive NLP research [63, 64, 65, 66, 67, 68, 17], there had been minimal user-centric analysis of SAsE prior to our work. We aimed to understand the impact of empirical disparities on SAsE speakers, whether this causes language adaptation when interacting with technology, and whether SAsE speakers desire better dialect support in language technology. These questions identify whether my proposed thesis direction addresses real user

[1]Full results for these further experiments are found in Appendix **??**

needs and wants.

2.2.1 Survey Design and Sampling

Our survey aims to (1) quantitatively assess language technology failure differences between SAsE and SAE speakers, and (2) gather qualitative feedback on user experiences and adaptations to understand if failures correspond to dialect usage. Respondents were informed that the study's purpose was "to understand how people use language to interact with technology." The survey begins with closed-ended questions establishing technology failure occurrences and types in English, followed by open-ended questions exploring user perceptions and adaptations.

Prolific was used to run this survey due to its large and diverse participant pool, high data quality [69, 70], 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 SAE speakers by pre-screening for US-born participants who only speak English.

The survey included 110 likely SAsE and 150 likely SAE speakers. We refined our pre-screened candidates using self-reported dialect information and shibboleth terms [71] distinguishing SAsE and SAE (*eggplant/brinjol, lentils/daal, elevator/lift*). For the SAE group, we excluded respondents who self-identified with other dialects or gave any SAsE-aligned answers. The SAsE group included only those who both self-identified with SAsE sub-dialects and provided SAsE-aligned responses.

2.2.2 Quantitative Survey Insights

Our survey results (see Figure 2.2) show that a majority of both SAsE (75%) and SAE (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 Assis-

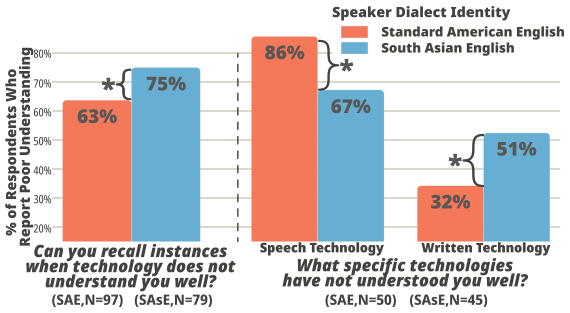

Figure 2.2: **Survey responses** to the questions *"Can you recall instances when technology does not understand you well?"* and *"What specific technologies have not understood you well?"*. * denotes significance at $P < 0.05$ using a Barnard Exact test.

Challenge	Example Keywords	Freq.
#1 Failures with stand-alone words	phrases, jargon, expressions, slang	43%
#2 Failures when switching between languages	foreign, local language, bilingual	18%
#3 Failures with colloquial dialect features	dialect, proper, standard, colloquial	20%

Table 2.3: **Reported challenges, corresponding keywords, and frequency of occurrence** in responses to open-ended questions.

tants 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 on text-based NLP **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 [72]. 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.

2.2.3 Qualitative Survey Insights

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) when **switching between languages**, and (3) with **dialect features**.

These identified challenges are not particularly surprising, given that they broadly cover the expected axes of variation in an Indigenized L2 variety of English. However, when we analyzed the frequency with which users cite each challenge (shown in Table 2.3) we found that the challenge most frequently cited by users (failures with stand-alone dialect words) diverge from those challenges emphasized in existing research (i.e. syntactic failures [17], switching between languages [26]). This points to a gap in current NLU research in addressing the wants of dialect speakers.

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

2.2.4 Constructing Corresponding Intrinsic Benchmarks

While some survey respondents mention extremely recent services like ChatGPT, most reference widely adopted technologies like customer service chatbots, search engines, and translation software. However, they don't cover all reported challenge categories, notably omitting stand-alone lexical variation—the largest issue mentioned by respondents. To better connect survey results with current state-of-the-art research, we curated new benchmarks to assess how respondent-reported variation affects LLMs.

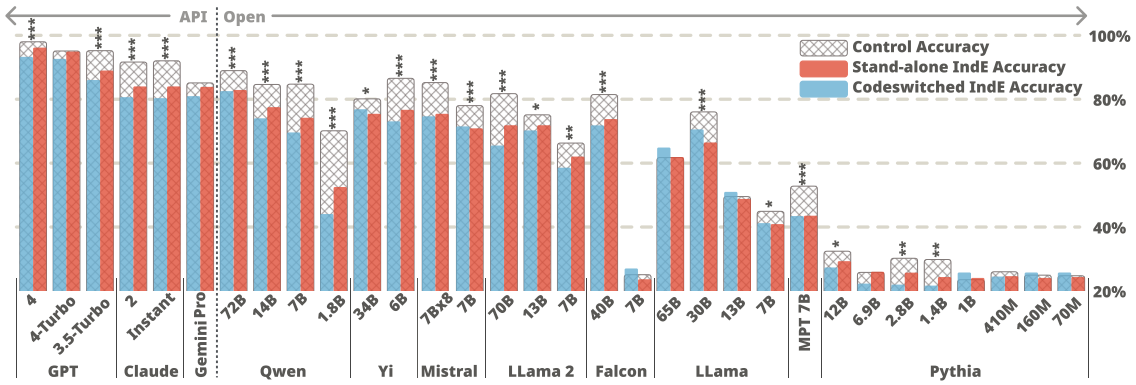

Figure 2.3: **Results for Wiktionary Benchmarks of both SAsE and Unmarked Lexical Knowledge**. *, **, and *** denote cases where overall performance is worse at P¡0.05, P¡0.01, and $P < 0.001$ respectively by a Bootstrap test. Control accuracy is for terms without any regional affiliation on Wiktionary.

As an intrinsic assessment of lexical understanding, we scraped 724 stand-alone terms and 317 loanwords from other South Asian languages from Wiktionary [73, 74] and formulated these as multiple choice questions. To assess syntactic understanding in isolation, we created a minimal pair syntactic language modeling evaluation in the style of Warstadt *et al.* [75] with 110 sentences aligned between SAE and Indian English [65]. Full details for dataset construction are presented in Appendix 6.2

We evaluate 8 series of open-source language models across both of these benchmarks. We evaluate an additional 3 industrial LLM providers on the lexical benchmarks, but are unable to evaluate them on the syntactic benchmark due to reliance on raw language modeling probabilities which API based models do not offer. LLMs demonstrate significant performance disparities on both SAsE benchmarks.

For lexical knowledge (shown in Figure 2.3), 14/15 open-access models with >60% control accuracy exhibit significant deficiencies ($P < 0.05$) on SAsE tasks. Though industrial models like GPT-4 achieve >90% accuracy, residual errors predominantly involve historical terminology, slurs, non-standard transliterations, and domain-specific lexicons. Notably, Indian English performance correlates strongly with control performance (p=0.98).

Regarding syntactic processing (shown in Figure 2.4), all evaluated models demonstrate near-perfect performance on SAE syntax while exhibiting statistically significant

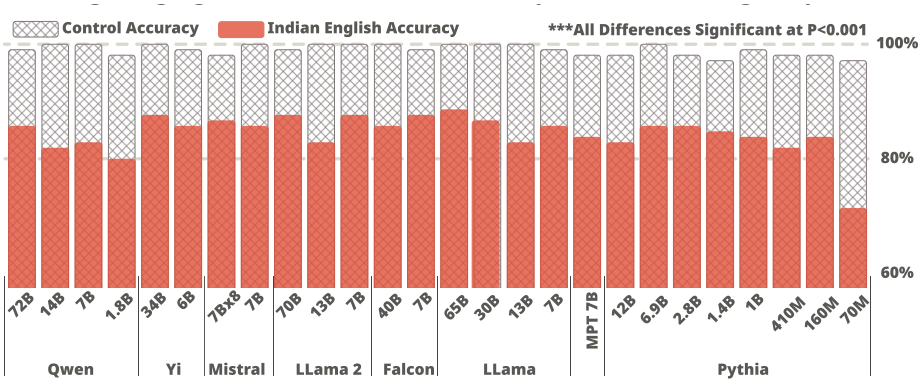

Figure 2.4: **Results for Minimal Pair Benchmark of both Indian and SAmE Syntactic Knowledge**. While the smallest models consistently perform nearly perfectly on the SAmE control, even the largest models perform significantly ($P < 0.001$) worse on the Indian English evaluation. Significance computed using a Bootstrap significance test.

degradation ($P < 0.001$) on SAsE syntax. Even the highest-performing model (LLama 65B) achieves only 89% accuracy. Despite this syntactic variation appears less frequently in user-reported challenges, this may reflect findings that syntactic understanding is less non-essential for functional NLP applications [76].

CHAPTER 3

METHODS FOR RAPID DIALECT ADAPTATION THROUGH FINETUNING

Building on the established evidence of dialect disparities in NLP systems and their impact on user experiences, this chapter introduces task-agnostic methods for rapid dialect adaptation that overcome limitations of previous approaches requiring costly task-specific annotations or data augmentation. The methods presented leverage alignment losses to optimize demographic parity at the representation level, tracing the evolution from initial optimization techniques to the full Task Agnostic Dialect Adapters (TADA) framework.

3.1 Distributional Alignment for Dialectal Parity

My methods work on dialectal robustness build on concepts from algorithmic fairness more broadly. However, since fairness is an abstract concept, rather than a mathematical construct, I begin this section by first defining and justifying the definition of robustness that I have pursued in my research.

3.1.1 Algorithmic Fairness Definitions

At their core, all definitions of algorithmic fairness involve making assumptions about the independence of a predictor $\hat{Y}$ and a demographic Z over which we would like to guarantee some definition of fairness. A significant number of these definitions such as test-fairness (predictions for all groups should be well calibrated) [77], equality of opportunity (the false positive rate should be independent of Z) [78] , and equalized odds(that odds of misclassifications should be equal across groups) [79] additionally make assumptions reliant on the true label Y.

In the context of Pretrained Language Models (PLMs), this reliance on the true label Y presents a significant problem. Firstly, even in a single class setting, for $Y \in \mathcal{Y}$ the com-

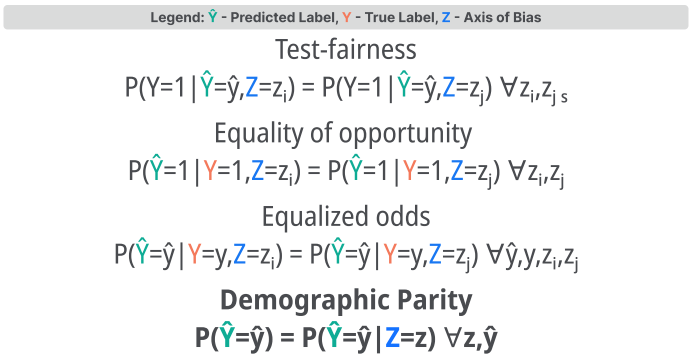

Figure 3.1: **Common formal definitions of Algorithmic Fairness** in general. My work has largely focused on pursuing demographic parity as it is the only definition which can be optimized in a task-agnostic fashion.

plexity of guaranteeing any of these definitions of fairness is dependent on $|\mathcal{Y}|$. For NLP, where generative and parsing tasks are common, we frequently have high-dimensional label spaces making these definitions of fairness incredibly difficult to guarantee. Perhaps more importantly, as PLMs are ideally used for many tasks, including some that are unknown at training time, it is intractable to guarantee fairness across *all* of these label spaces.

Instead, my research focuses on a definition of fairness that is tractable to optimize in a task-agnostic fashion: demographic parity. A predictor satisfies demographic parity if only $\hat{Y}$ and Z are independent — more formally: $P(\hat{Y} = \hat{y}) = P(\hat{Y} = \hat{y}|Z = z) \quad \forall z, \forall \hat{y}$. While this definition is still dependent on $\hat{Y}$, this can easily be addressed in a task-agnostic fashion by instead optimizing for demographic parity on the final hidden dimension h of a PLM. Debiasing h is sufficient because if the final hidden representation is independent of the protected attribute Z, then any downstream task that uses this representation as input will inherit this independence, thereby satisfying demographic parity regardless of the specific prediction task without further training.

3.1.2 Constrained Adversarial Optimization

I first explored this technique for improving the robustness of PLMs throughout finetuning in Held *et al.* [26] drawing inspiration from more general work in algorithmic fairness

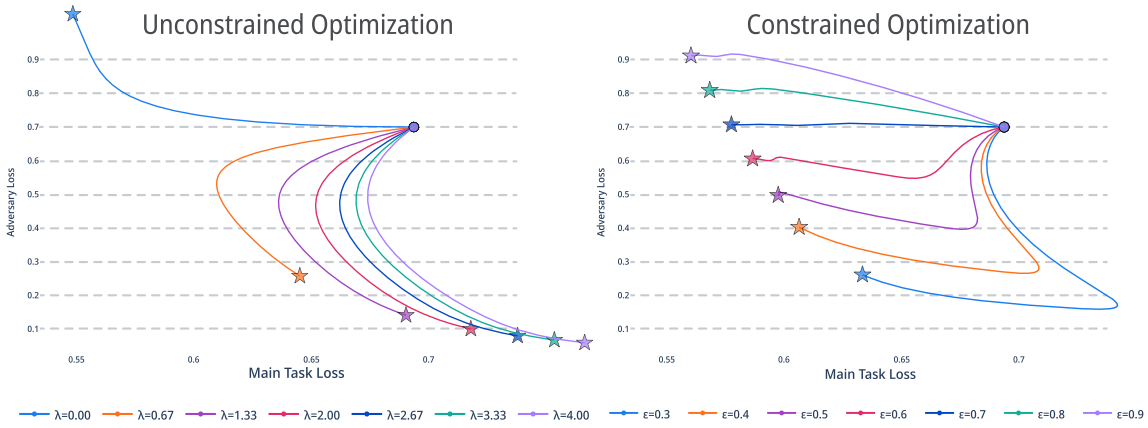

Figure 3.2: **Comparison of Optimization Approaches** for adversarial alignment. While unconstrained optimization is unpredictable with respect to λ, constrained optimization allows selecting a semantically meaningful ϵ in advance.

using adversarial methods [80, 81]. Beyond fairness, using adversarial learning to remove undesirable features had been discovered and applied separately in transfer learning [82] and privacy preservation [83]. As shown in Figure 3.4, this can be used to align dialects by training a critic model to distinguish SAE data and data in other dialects. The critic is trained to identify the dialect of the input based on the final hidden state, while the main model is trained to make the dialects indistinguishable.

As I have highlighted, on its own this method is well established. However, as noted in [81] work "adversarial training method is hard to get right" and "getting the hyperparameters wrong results in quick divergence" which undermines the theoretical elegance of adversarial methods as a simple way to optimize fairness.

To understand why this is a difficult optimization problem, we can look at the dynamics of the Min-Max game between the adversary and the model we intend to debias. Demographic parity is achieved only when the adversaries loss is equal to the entropy of Z. Naively, the loss can increase beyond the entropy of Z when the adversary and the main model are mismatched in learning capacity. This leads to instability in the training procedure as the adversary can simply invert the sign of the predictions to achieve a better loss. Ultimately, this oscillation is what makes the adversarial learning procedure difficult to fit

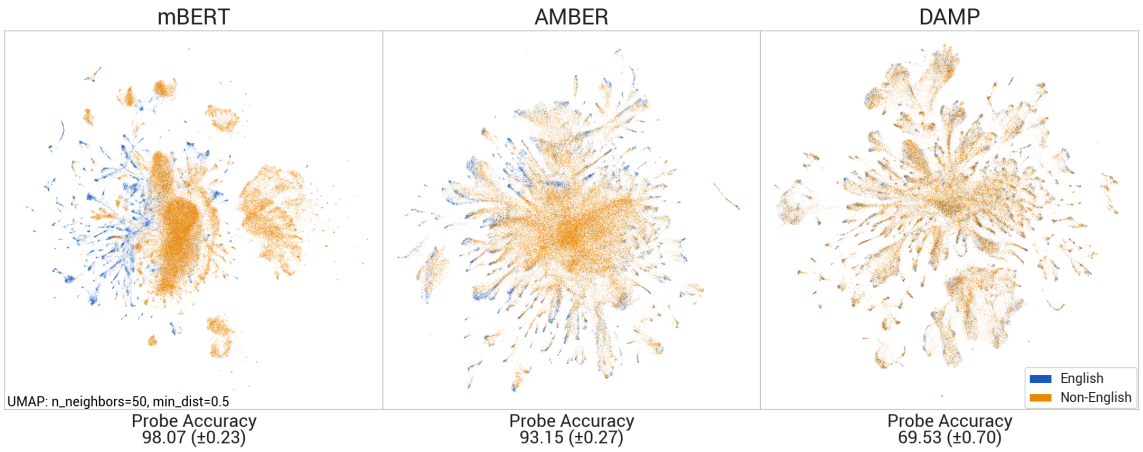

Figure 3.3: **Codeswitching alignment results** from [26] showing that constrained adversarial optimization leads to more strongly aligned representations across languages with both visual checks and adversarial probing results. Throughout my work, I have used alignment as a fairness objective.

since many hyperparameter configurations lead to mode collapse.

If we instead view the entropy of Z as a constraint on the adversaries loss, the optimization procedure can be vastly simplified. Rather than an unconstrained gradient ascent for the adversary, we optimize within the space of algorithmically fair solutions with minimal additional computation cost using the differential method of multipliers [84]. While originally introduced in my work specific to task-oriented parsing during an internship at Google, the method itself is not parsing specific and therefore I quickly leveraged it in pursuit of the larger research goal of task-agnostic robustness.

3.2 Task-Agnostic Dialect Adaptation

The optimization improvements above largely became useful in my thesis direction through the development of Task-Agnostic Dialect Adapters (TADA). This approach draws from successes in multilingual transfer learning, where domain adaptation using small parallel datasets has proven competitive with large-scale augmentation [85]. TADA extends this concept to dialect adaptation with both simple alignment loss and a continuation of the adversarial debiasing methods above. Overall, our goal is to remove dialect specific infor-

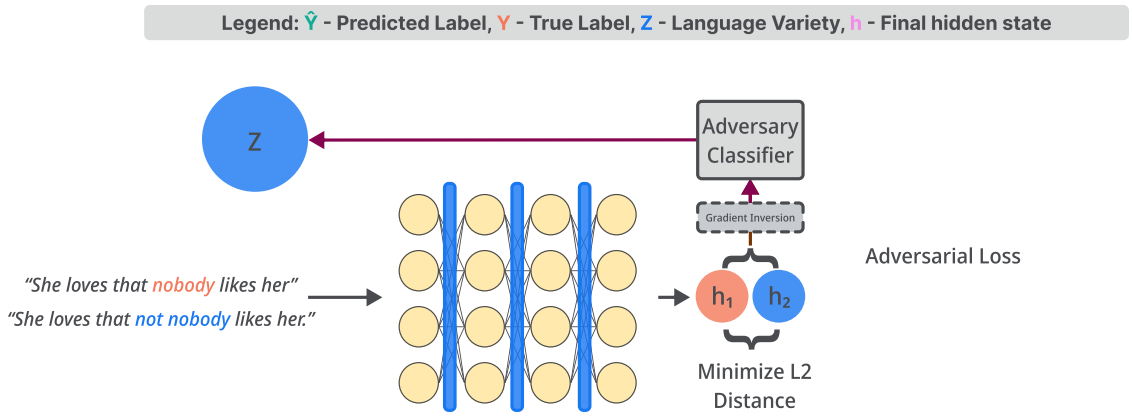

Figure 3.4: **Task-Agnostic Adapter training flow** with both sequence and token level alignment loss between SAE and a target dialect. When stacked before task-specific SAE adapters, TADA provides dialect robustness for the target task.

mation from the final hidden states of a PLM in a task-agnostic fashion in order to improve the performance of models on non-SAE dialects in a plug-and-play fashion.

3.2.1 Loss Function and Data Construction

Without a task-loss, it is likely that the early gradients from the adversarial methods would be largely meaningless. Therefore, we first provide a smooth loss signal by simply minimizing the distance of a pooled representation of frozen SAE inputs and learnable non-SAE inputs:

$$L_{seq} = |C\vec{L}S_{sae} - C\vec{L}S_{dial}|_2 \tag{3.1}$$

However, this leaves significant room for dialect disparity to continue to exist in token-level representations, we leverage the optimization method from the previous section to optimize for token level parity using a transformer-based adversary [86]. Given an adversarial scoring network Adv, frozen SAE representation $\vec{SAE}$, and post-TADA non-SAE representation $\vec{Dial}$, we define this token-level loss:

$$L_{tok} = -\mathrm{Adv}(\vec{Dial}) \tag{3.2}$$

TADA utilizes only 1,000 synthetic sentence-parallel examples generated via rule-based transformations from Multi-VALUE [17]. This intentional data limitation means that the data could feasibly be replaced with human-translated examples [65] and enables adaptation to other language varieties where small parallel corpora exist but systems like Multi-VALUE do not. Using these, we train invertible adapters [87] to minimize the combined loss $L_{TADA} = L_{seq} + L_{tok}$. The full training procedure is shown in Figure 3.4.

The benefit of these adapters, trained on alignment, is that they can be composed with *task-specific* adapters at test time. Without any further training, this allows TADA modules to provide the benefits of dialect robustness to a specific task without any further training.

3.2.2 Evaluating Trained Adapters

Since ours is the first work to attempt task-agnostic dialect adaptation, we benchmark TADA in comparison to prior task-specific methods in Table 3.1.

We first establish pure SAE baselines for both full finetuning and adapter training [88]. Interestingly, the gap between SAE performance and AAE performance is similar for adapters (-8.8) and full finetuning (-8.9) when trained on SAE. The minimal effects of the limited capacity of adapters on disparity indicate that dialectal discrepancy is largely within the pretrained LLM before finetuning. Without mitigation, SAE models alone perform poorly on non-SAE input.

We then train two task-specific dialect mitigation following the approach of VALUE, which augments training data with pseudo-dialect examples during finetuning. This is a strong baseline, as it allows the model to adapt specifically to in-domain augmented examples rather than the general sentences used to align TADA modules. When trained on augmented data, adapters (80.8 Avg.) seem to outperform full finetuning (77.5 Avg.).

Finally, we combine TADA with task-specific SAE modules for our task-agnostic approach. TADA succeeds in our goal of generalizable performance improvements, yielding improved robustness for 6 out of 7 tasks for an average increase of 2.8 points on the GLUE

Dialect Adaptation Details			AAVE GLUE Perf.
Approach	Method	Dialect Params.	Mean
N/A	Finetuning	0	75.1
N/A	Adapters	0	74.7
VALUE	Finetuning	$T \times 110M$	77.5
VALUE	Adapters	$T \times 895K$	80.8
TADA	Adapters	$895K$	77.5+

Table 3.1: **AAVE Adaptation results** of RoBERTa Base [53]. T is the number of target tasks for dialect adaptation. Tasks where TADA improves the performance of task-specific SAE adapters, are marked with +.

	Mean	
Test Dialect	Orig.	TADA
SAE	83.5	83.5
AAVE	74.7	77.5 (+2.8)
Indian	74.4	74.7 (+0.3)
Nigerian	76.3	76.7 (+0.4)
Singapore	70.9	74.8 (+3.9)

Figure 3.5: **Multi-Dialectal** evaluation results (Mean across all tasks) for 4 Non-SAE Dialect Variants of GLUE created using Multi-VALUE.

benchmark. However, TADA performs 4% worse on average than task-specific VALUE-augmented adapters. These adapters are trained on larger amounts of dialectal training data directly from each task than TADA, which likely explains their superiority. However, as noted in the table these approaches scale training and storage linearly with the number of tasks, while TADA requires only a constant overhead.

We then test whether TADA generalizes across regional dialects using 3 global dialects in addition to AAVE in Table 3.5. TADA improves performance for African American (+2.8), Indian (+0.3), Nigerian (+0.4), and Singaporean (+3.9) Englishes respectively. These results demonstrate TADA's potential as a general tool for dialect adaptation, both across dialects and across tasks. However, a notable limitation of TADA is that it still relies on dialect-specific training which I aimed to address in follow-up work led by mentees.

3.2.3 Further Work Incorporating Linguistic Knowledge into TADA

Following the development of TADA, I advised two extensions focused on novel neural architectures that incorporate linguistic knowledge into the learning process. Both approaches significantly improved data efficiency and generalization capabilities.

DADA [28] moved beyond dialect-level adaptation to a more fine-grained approach working at the level of individual linguistic features. While TADA required separate adapters for each dialect, DADA introduced a modular architecture that captures specific

linguistic features through individual adapters that can be dynamically composed. This eliminated the need for dialect identification systems by focusing on the linguistic features present in the input, regardless of their classification into traditional dialect categories.

DADA trained nearly 200 feature adapters, each capturing a specific linguistic transformation rule. The compositional architecture enabled both targeted adaptation to specific dialect variants and simultaneous adaptation to various dialects by leveraging their feature commonalities. Experiments across five English dialects (AppE, ChcE, CollSgE, IndE, AAVE) demonstrated DADA's effectiveness for both single-task models and instruction-tuned language models, while also exhibiting strong interpretability through adapter activation patterns.

HyperLoRA [29] addressed a more fundamental challenge: adapting to completely unseen dialects without any dialect-specific training data. This approach leveraged expert linguistic knowledge in the form of typological feature vectors from dialectology research. A hypernetwork architecture was developed to generate Low-Rank Adaptation (LoRA) parameters conditioned on these linguistic feature vectors, disentangling dialect-specific and cross-dialectal information and improving generalization in a task-agnostic fashion.

HyperLoRA achieved competitive performance across multiple unseen dialects without requiring any dialect-specific annotations, demonstrating that expert knowledge could effectively substitute for approximately 250 dialectal annotations per dialect. This marked a significant advance in resource efficiency and scalability.

Together, these extensions addressed the primary limitations identified in the initial TADA work, providing more flexible, efficient, and scalable approaches to dialect adaptation that can work with evolving dialect landscapes and limited resources.

CHAPTER 4

FORECASTING DIALECT ROBUSTNESS AS A FUNCTION OF SCALE

While the prior chapter introduced plug-and-play adaptation modules for existing models, such methods require individual practitioners to proactively seek out dialect robustness interventions. In practice, most teams are unlikely to prioritize robustness when building real-world applications, especially as PLMs are increasingly deployed as general-purpose components in software where robustness is rarely tested systematically [24]. Simultaneously, due to the tremendous zero-shot capabilities which seem to emerge from pretraining [89], it is not unreasonable to expect linguistic robustness may similarly emerge.

This belief in the unreasonable effectiveness of scale became increasingly prevalent after the launch of ChatGPT and thereby motivated a natural critical question of my line of work: do we need to embed dialect robustness into LLMs explicitly or will scale solve the challenge? Naturally, since much of my work focuses on studying targeted metrics, this question raises major implications for the long-term meaning of my work. As such, I aimed to test this question empirically in Held *et al.* [90] which is covered in this section.

4.1 Scaling Laws and Dialect Performance

Classical scaling laws describe loss L as a power-law function of compute F (in FLOPs) [30]:

$$L(F) = \alpha F^{-\beta},$$

where $\alpha > 0$ denotes the initial loss level and $\beta > 0$ the rate of improvement with scale. These constants are typically estimated empirically. While these laws have been validated primarily on in-distribution data, their behavior across linguistic subpopulations remains poorly understood.

4.1.1 Relative Scaling Laws

Relative scaling laws extend the classical power-law formulation to quantify how performance gaps evolve with compute. If absolute error decreases as a power law, then the *relative error* between a baseline dialect (b) and a treatment dialect (t) can be expressed as

$$G(F) = \frac{E_t(F)}{E_b(F)} = \gamma F^{\Delta\beta},$$

where $\gamma = \alpha_t/\alpha_b$ captures the initial disparity and $\Delta\beta = \beta_b - \beta_t$ the difference in scaling rates. If $\Delta\beta < 0$, gaps narrow with scale; if $\Delta\beta > 0$, they widen; if $\Delta\beta = 0$, they remain constant.

This form parallels the subgroup laws[91], which predicts that effective compute decomposes into subgroup and shared components:

$$L_i(F_i, F_t) = \frac{B_i}{F_i^{\beta_i}} + \frac{C_t}{F_t^{\alpha_t}},$$

where F_i and F_t denote subgroup- and total-compute scales, respectively. If $\alpha_t > \beta_i$, subgroup losses converge with scale; if $\alpha_t < \beta_i$, disparities persist or widen.

Our formulation is looser — we do not require subgroup allocations — but the sign of $\Delta\beta$ still forecasts whether gaps shrink or persist. While relative loss can correspond to small absolute differences at low loss, small absolute loss gaps can lead to large differences in downstream utility for large scale models [92, 93] which motivates this scale-invariant metric rather than absolute disparity.

4.1.2 Hypothesis

While digital access has expanded over time [94]. Even as access to the internet has equalized, the historical digital divide has created a ”value lock” in the online digital archive [95]. This has set up the internet as a biased sample, with the majority of content coming from

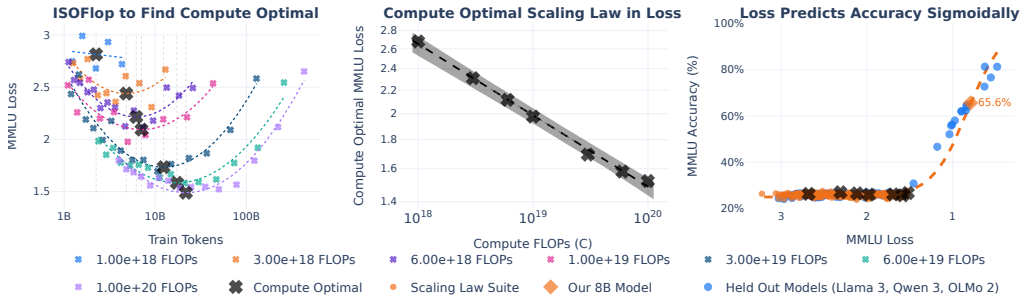

Figure 4.1: **Compute-optimal scaling and downstream forecasting. Left:** For each FLOP budget, we sweep token and model size to select the compute-optimal token count. **Middle:** Along these compute-optimal points, we estimate how task or subgroup loss scales as a function of compute. **Right:** We show this loss correlates tightly with accuracy sigmoidally, allowing loss to serve as a proxy for downstream progress while measuring effects at reduced scale.

North America and Europe [96, 97, 98]. In work not covered in this thesis, I covered these sociotechnical feedback loops in further depth[36].

As such, under the functional forms from Rolf *et al.* [91], we would expect scaling laws to be steeper for regions with earlier and more prevalent internet access, as the topics and linguistic patterns of these users have had a longer time to accumulate on the web.

However, since it is difficult to directly determine the provenance of each post on the web, we must make a few simplifying assumptions on our hypothesis. Firstly, we will use our simple power law form of relative scaling laws. Secondly, we will assume each user with internet access is equally likely to produce data which ends up online, regardless of their regional identity. Under this model, we would expect to see a correlation between our relative scaling law slopes and the number of internet users a region has.

4.1.3 Experimental Setup

We train models using the Qwen 3 architecture [99] under fixed compute (IsoFLOP) budgets ranging from 10^{18} to 10^{20} FLOPs. While IsoFLOPs are not strictly necessary for scaling laws, prior work [100, 101, 102] has argued that the IsoFLOP-based approach is more stable and therefore less exposed to reproducibility issues than alternative formulations.

Scaling models should be trained such that performance variance is primarily explained

by compute, model size, and data size. Without consistent hyperparameter tuning, scaling outcomes can be meaningfully confounded [103]. Since a full grid search is infeasible, we generalize a tuned configuration [104] using heuristic reparameterizations.

Our approach follows two principles: (i) hyperparameters should be explicit functions of model width and FLOP budget; and (ii) training should be stable across runs, since instabilities such as loss spikes would introduce noise into scaling comparisons. I describe this in more depth in Appendix 6.3.

We train models with the same configuration across three datasets to reflect different pretraining data distributions. COMMONPILE [105] includes only permissively licensed data, downsampling non-permissive web sources in favor of public domain and openly licensed material. In contrast, the DCLM BASELINE [106] is drawn entirely from web crawl data but filtered and deduplicated to isolate a high-quality subset. Finally, NEMOTRON-CC [107] combines large-scale real web data with synthetic rephrasings, representing a hybrid of natural and synthetic text. Comparing scaling behavior across these settings enables assessments of the role of training data in relative scaling results.

4.1.4 Empirical Analysis of Dialect Scaling

We then evaluate compute optimal models with the International Corpus of English (ICE) [108], which includes $\sim$1M words per region spanning spoken and written registers under consistent national sampling from speakers with high-school or higher levels of education.

Figure 4.2 shows that absolute performance rises for all dialects, yet gaps with U.S. English persist or even grow for some dialects. Across training corpora, disparity vs. U.S. English decreases for Canada ($+0.3$–0.5% per $10\times$ FLOPs), is roughly flat for Singapore (0–0.1%), and increases for Sri Lanka (-0.5–-0.9%) and Nigeria (-0.3–-0.8%). Even regions with similar initial accuracy can diverge: for the CommonPile, Nigeria and Singapore start within one point of eachother, but by 10^{20} FLOPs Singapore is $\approx -2\%$ while Nigeria is $\approx -5\%$.

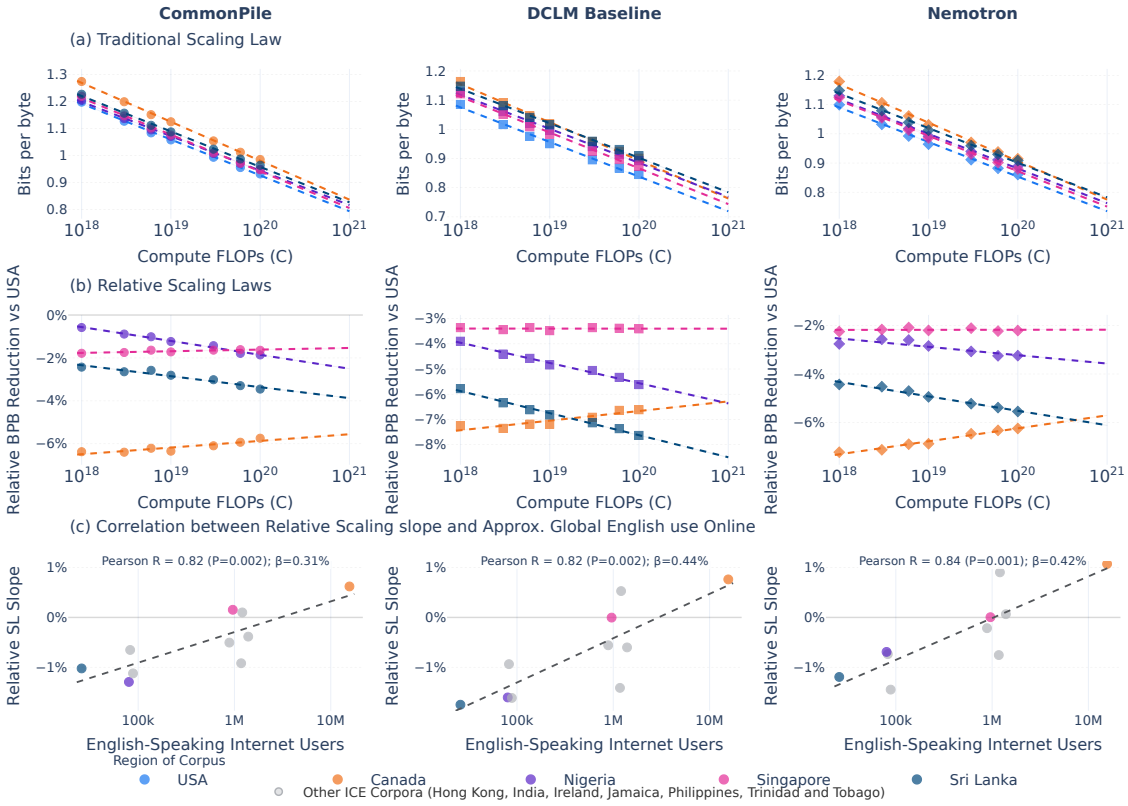

Figure 4.2: **Relative scaling of written Global Englishes.** (a) Absolute bits-per-byte (BPB) vs. compute. (b) Performance relative to U.S. English (dashed). (c) Relative slopes vs. English-speaking internet prevalence (ICE era).

These patterns lead to unstable orderings, so today's lowest-performing regions may not be the most urgent under scaling. For the CommonPile, Singapore begins below Nigeria but crosses at 6×10^{19} FLOPs; in DCLM, Canada and Sri Lanka cross at 3×10^{19} FLOPs. These shifts are overlooked by point estimates, highlighting the importance of modeling scaling trends for forecasting subgroup disparities.

We also find evidence in support of [91], which hypothesizes that subgroup representation in training data primarily affects scaling terms. While the BPB *intercepts* show no clear correlation with prevalence, countries with larger estimated online English-speaking populations — such as Canada and Singapore — have neutral or positive relative scaling *slopes* and those with smaller populations at the time ICE was collected — such as Sri Lanka and Nigeria — have negative relative scaling slopes. Across all 10 ICE corpora, slope–prevalence correlation is robust across training datasets (Pearson $R = 0.82$–0.84,

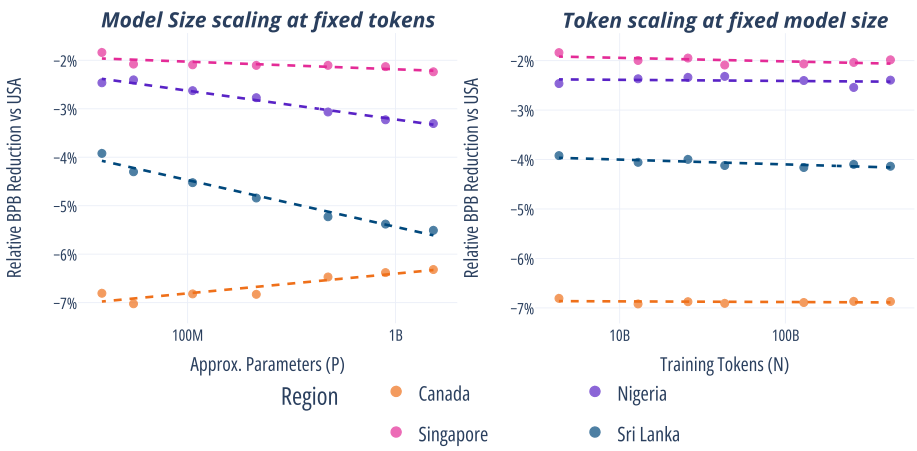

Figure 4.3: **Isolating scaling factors.** Varying capacity allocation at fixed FLOPs shifts relative performance; varying token count at fixed compute largely preserves the ordering.

$p < 0.005$), corresponding to a 0.3–0.4% relative error slope improvement per ten-fold increase in speaker population.

In contrast to our compute-optimal results, the minimal prior work studying how scale impacts robustness to language variation [109] looked only at parameter scaling in isolation. We revisit these analyses in Figure 4.3, evaluating relative scaling laws for both parameter and data scaling in isolation. When scaling model size at a fixed 10B-token budget, relative performance shifts similarly to compute-optimal scaling. By contrast, when scaling training tokens at a fixed architecture, the lines remain almost perfectly parallel: all regions improve together, but their ordering relative to U.S. English is unchanged. This indicates that model-size scaling drives the observed shifts in relative performance, while data scaling leaves relative performance largely unchanged.

Across ICE corpora, relative slope correlates strongly with the size of the online English-speaking population (Pearson $R \approx 0.82$–0.84, $p < 0.005$): larger-population dialects have neutral or positive slopes, while smaller-population dialects have negative slopes. These patterns are consistent with subgroup scaling predictions in which underrepresented subpopulations experience diminishing returns with scale [91].

Summary of Findings. Scaling improves average performance across dialects but is not a universal equalizer. Some gaps shrink while others widen, and aggregate gains can mask systematic disparities. Capacity—not data volume—primarily drives relative shifts, favoring dialects with stronger representation.

4.2 Discussion and Implications

Our results provide the first empirical characterization of dialectal scaling dynamics during pretraining. Framing dialect robustness in the language of scaling laws enables two advances. First, it yields *forecasts*: given current mixtures, we can anticipate which dialect gaps are likely to persist with additional compute. Second, it provides a common currency to compare the compute-efficiency of targeted interventions—synthetic augmentation (Multi-VALUE [17]), task-agnostic adaptation (TADA [27]), and multi-domain data curation [110])—against the baseline of "just scale it."

Conceptually, these findings caution against assuming that scale alone will equalize performance across English varieties. In fact, capacity-driven gains can preferentially accrue to well-represented dialects, risking widened disparities absent explicit countermeasures. Practically, this argues for deliberate pretraining design—balanced data, subgroup-aware objectives—and post hoc adaptation to ensure equitable performance. The dialect scaling framework thus links methodological contributions from earlier chapters to the foundation-model era: robust, inclusive language technology requires more than scale; it requires *design* for linguistic diversity.

CHAPTER 5
CONCLUSION

This dissertation examines how to evaluate, forecast, and improve the robustness of NLP systems in the face of linguistic variation within English. The central thesis is straightforward: English dialect diversity is systematic enough to model, yet underrepresented enough to produce measurable and meaningful disparities. This systematicity means that interventions can be implemented which give both theoretically and empirically strong robustness to dialectal variation. Due to social biases in web data, simple scaling does not resolve these robustness issues and, in the worst case, further entrenches them.

5.1 Summary of Contributions

Measuring dialect disparities (Ch. 2). I combined controlled, causal evaluation with user-centered evidence. The Multi-VALUE toolkit extends feature-controlled counterfactual tests to 50 global Englishes, making it possible to attribute performance differences to concrete linguistic features. A survey of South Asian English (SAsE) speakers connects these quantitative results to practical experience: participants describe distinct failure modes, report adapting their language for better system performance, and express a consistent preference for dialect-aware technology.

Rapid, task-agnostic adaptation (Ch. 3). I introduced *Task-Agnostic Dialect Adapters* (TADA), which improve robustness by aligning SAE and non-SAE representations through sequence- and token-level objectives. TADA composes with task adapters at inference, providing plug-and-play gains without retraining. Follow-up work expanded this line: DADA decomposes adaptation into interpretable linguistic modules, and HyperLoRA conditions adapters on typological vectors to generalize to unseen dialects.

Dialect dynamics under scaling (Ch. 4). I analyzed how dialect performance changes as models scale. While all dialects benefit in absolute accuracy, relative disparities often persist or widen. Model capacity—not training token count—emerges as the main driver of divergence. Dialects that are better represented in pretraining data gain more from scale, matching theoretical predictions for minority subpopulations. Viewing fairness through scaling laws allows data- and compute-aware forecasts of where inequities will remain unless addressed.

5.2 Closing Remarks

English is a shared language with many centers of gravity. Systems tuned to dominant varieties risk overlooking large and consequential parts of that landscape. While much of my own work is motivated in the lens of more fair machine learning systems, there are meaningful industrial reasons to optimize for robustness to global English variation. Many nations, such as India and Nigeria, have rapidly growing populations of English speakers and as the digital divide closes, these communities are increasingly online. Due to the imbalances of historical web data, our scaling law analysis shows that these communities may not be well served by simply scaling up existing approaches, leaving room for applications which better intervene on dialectal representation to capture significant market share.

My dissertation has build out the tools for a more intentional approach to building linguistically robust NLP systems, providing open-source tools to evaluate robustness systematically, adapt foundation models efficiently, and generally treat dialect representation as a core design concern. By focusing on methods which are generally task-agnostic, these methods are amenable with the modern foundation model paradigm and allow practitioners to intervene in a cost-efficient way while still enabling multi-task transfer. While naive scaling on the web is unlikely to provide robust models, I firmly believe that inentional application of the robustness methods developed in this thesis can build systems which more fluidly adapt to serve an evergrowing population of potential users of LLMs.

REFERENCES

[1] E. M. Bender, "On achieving and evaluating language-independence in nlp," *Linguistic Issues in Language Technology*, vol. 6, 2011.

[2] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (technology) is power: A critical survey of "bias" in nlp," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 5454–5476.

[3] D. Hovy and D. Yang, "The importance of modeling social factors of language: Theory and practice," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova *et al.*, Eds., Association for Computational Linguistics, Jun. 2021, pp. 588–602.

[4] D. Hershcovich *et al.*, "Challenges and strategies in cross-cultural NLP," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6997–7013.

[5] P. W. Koh *et al.*, "Wilds: A benchmark of in-the-wild distribution shifts," in *International conference on machine learning*, PMLR, 2021, pp. 5637–5664.

[6] C. Lehmann, "Grammaticalization: Synchronic variation and diachronic change," *Lingua e stile*, vol. 20, no. 3, pp. 303–318, 1985.

[7] J. K. Chambers and P. Trudgill, *Dialectology*. Cambridge University Press, 1998.

[8] W. Labov, "The intersection of sex and social class in the course of linguistic change," *Language variation and change*, vol. 2, no. 2, pp. 205–254, 1990.

[9] J. R. Rickford, *African American Vernacular English: Features, Evolution, Educational Implications*. Malden, MA: Wiley-Blackwell, 1999, ISBN: 9780631212454.

[10] D. Biber, *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press, 1995.

[11] P. Eckert, "Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation," *Annual review of Anthropology*, vol. 41, no. 1, pp. 87–100, 2012.

[12] D. Sharma, *From deficit to dialect: The evolution of English in India and Singapore*. Oxford University Press, 2023.

[13] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The state and fate of linguistic diversity and inclusion in the nlp world," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6282–6293.

[14] D. Jurgens, Y. Tsvetkov, and D. Jurafsky, "Writer profiling without the writer's text," in *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part II 9*, Springer, 2017, pp. 537–558.

[15] B. B. Kachru, "World englishes: Approaches, issues and resources," *Language teaching*, vol. 25, no. 1, pp. 1–14, 1992.

[16] S. Bird, "Local languages, third spaces, and other high-resource scenarios," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 7817–7829.

[17] W. Held, C. Ziems, J. Yang, J. Dhamala, R. Gupta, and D. Yang, "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)*, 2023, pp. 744–768.

[18] C. Ziems, J. Chen, C. Harris, J. Anderson, and D. Yang, "Value: Understanding dialect disparity in nlu," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 3701–3720.

[19] W. Held, F. Holt, and D. Yang, "Perceptions of language technology failures from south asian english speakers," in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 4067–4081.

[20] S. L. Blodgett, J. Wei, and B. O'Connor, "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)*, 2018, pp. 1415–1425.

[21] T. Blevins, R. Kwiatkowski, J. C. Macbeth, K. McKeown, D. Patton, and O. Rambow, "Automatically processing tweets from gang-involved youth: Towards detecting loss and aggression," 2016.

[22] A. Jørgensen, D. Hovy, A. Søgaard, *et al.*, "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, 2016.

[23] D. Jurgens, Y. Tsvetkov, and D. Jurafsky, "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)*, R. Barzilay and M.-Y. Kan, Eds., Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 51–57.

[24] Q. Yang, J. Suh, N.-C. Chen, and G. Ramos, "Grounding interactive machine learning tool design in how non-experts actually build models," in *Proceedings of the 2018 Designing Interactive Systems Conference*, ser. DIS '18, Hong Kong, China: Association for Computing Machinery, 2018, pp. 573–584, ISBN: 9781450351980.

[25] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi, "Data decisions and theoretical implications when adversarially learning fair representations," *arXiv preprint arXiv:1707.00075*, 2017.

[26] W. Held *et al.*, "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)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 3586–3604.

[27] W. Held, C. Ziems, and D. Yang, "TADA : Task agnostic dialect adapters for English," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 813–824.

[28] Y. Liu, W. Held, and D. Yang, "DADA: Dialect adaptation via dynamic aggregation of linguistic rules," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 13 776–13 793.

[29] C. Lv *et al.*, "HyperLoRA: Efficient cross-task generalization via constrained low-rank adapters generation," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 16 376–16 393.

[30] J. Kaplan *et al.*, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[31] J. Hoffmann *et al.*, "Training compute-optimal large language models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022, pp. 30 016–30 030.

[32] T. Eloundou, S. Manning, P. Mishkin, and D. Rock, "Gpts are gpts: Labor market impact potential of llms," *Science*, vol. 384, no. 6702, pp. 1306–1308, 2024.

[33] K. Handa *et al.*, *Which economic tasks are performed with ai? evidence from millions of claude conversations*, 2025. arXiv: 2503.04761 [cs.CY].

[34] E. Brynjolfsson, D. Li, and L. Raymond, *Generative ai at work*, 2024. arXiv: 2304.11771 [econ.GN].

[35] D. Acemoglu and P. Restrepo, "Artificial intelligence, automation, and work," in *The economics of artificial intelligence: An agenda*, University of Chicago Press, 2018, pp. 197–236.

[36] W. Held, C. Harris, M. Best, and D. Yang, "A material lens on coloniality in nlp," *arXiv preprint arXiv:2311.08391*, 2023.

[37] M. J. Ryan, W. Held, and D. Yang, "Unintended impacts of LLM alignment on global representation," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 16 121–16 140.

[38] J. Myung *et al.*, "Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages," in *Advances in Neural Information Processing Systems*, A. Globerson *et al.*, Eds., vol. 37, Curran Associates, Inc., 2024, pp. 78 104–78 146.

[39] T. Davidson, D. Bhattacharya, and I. Weber, "Racial bias in hate speech and abusive language detection datasets," in *Proceedings of the Third Workshop on Abusive Language Online*, Florence, Italy: Association for Computational Linguistics, 2019, pp. 25–35.

[40] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Hate speech detection and racial bias mitigation in social media based on bert model," *PloS one*, vol. 15, no. 8, e0237861, 2020.

[41] A. Rios, "Fuzze: Fuzzy fairness evaluation of offensive language classifiers on african-american english," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, AAAI Press, 2020, pp. 881–889.

[42] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, "The risk of racial bias in hate speech detection," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, 2019, pp. 1668–1678.

[43] X. Zhou, M. Sap, S. Swayamdipta, Y. Choi, and N. Smith, "Challenges in automated debiasing for toxic language detection," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online: Association for Computational Linguistics, 2021, pp. 3143–3155.

[44] B. Lwowski and A. Rios, "The risk of racial bias while tracking influenza-related content on social media using machine learning," *Journal of the American Medical Informatics Association*, vol. 28, no. 4, pp. 839–849, 2021.

[45] B. Kortmann, K. Lunkenheimer, and K. Ehret, Eds., *eWAVE*. 2020.

[46] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "Spacy: Industrial-strength natural language processing in python," 2020.

[47] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, *Stanza: A python natural language processing toolkit for many human languages*, 2020. arXiv: 2003.07082 [cs.CL].

[48] Z. Wu, A. Tamkin, and I. Papadimitriou, "Oolong: Investigating what makes transfer learning hard with controlled studies," in *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[49] D. Blasi, A. Anastasopoulos, and G. Neubig, "Systematic inequalities in language technology performance across the world's languages," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5486–5505.

[50] F. Faisal, S. Keshava, M. M. I. Alam, and A. Anastasopoulos, "SD-QA: Spoken dialectal question answering for the real world," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds., Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3296–3315.

[51] S. Reddy, D. Chen, and C. D. Manning, "CoQA: A conversational question answering challenge," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 249–266, 2019.

[52] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186.

[53] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *ArXiv preprint*, vol. abs/1907.11692, 2019.

[54] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[55] M. Lewis *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Association for Computational Linguistics, Jul. 2020, pp. 7871–7880.

[56] T. Yu *et al.*, "Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 3911–3921.

[57] M. R. Costa-jussà *et al.*, "No language left behind: Scaling human-centered machine translation," *ArXiv preprint*, vol. abs/2207.04672, 2022.

[58] L. Barrault *et al.*, "Findings of the 2019 conference on machine translation (WMT19)," in *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, O. Bojar *et al.*, Eds., Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 1–61.

[59] Z. Mengesha, C. Heldreth, M. Lahav, J. Sublewski, and E. Tuennerman, ""i don't think these devices are very culturally sensitive."—impact of automated speech recognition errors on african americans," *Frontiers in Artificial Intelligence*, vol. 4, p. 169, 2021.

[60] R. Gargesh, "South asian englishes," *The handbook of world Englishes*, pp. 105–134, 2019.

[61] A. F. Gupta, "Indian english," *The Handbook of World Englishes*, vol. 7, pp. 203–222, 2010.

[62] B. B. Kachru, "The indianness in indian english," *Word*, vol. 21, no. 3, pp. 391–410, 1965.

[63] A. Irvine, J. Weese, and C. Callison-Burch, "Processing informal, romanized pakistani text messages," in *Proceedings of the Second Workshop on Language in Social Media*, 2012, pp. 75–78.

[64] R. Sarkar, S. Mahinder, and A. KhudaBukhsh, "The non-native speaker aspect: Indian English in social media," in *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, W. Xu, A. Ritter, T. Baldwin, and A. Rahimi, Eds., Association for Computational Linguistics, Nov. 2020, pp. 61–70.

[65] D. Demszky, D. Sharma, J. Clark, V. Prabhakaran, and J. Eisenstein, "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*, Association for Computational Linguistics, Jun. 2021, pp. 2315–2338.

[66] T. Masis, A. Neal, L. Green, and B. O'Connor, "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*, O. Serikov *et al.*, Eds., Gyeongju, Republic of Korea: International Conference on Computational Linguistics, Oct. 2022, pp. 11–25.

[67] J. Sun *et al.*, "Dialect-robust evaluation of generated text," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 6010–6028.

[68] J. Eisenstein, V. Prabhakaran, C. Rivera, D. Demszky, and D. Sharma, "MD3: The Multi-Dialect Dataset of Dialogues," in *Proc. INTERSPEECH 2023*, 2023, pp. 4059–4063.

[69] P. Eyal, R. David, G. Andrew, E. Zak, and D. Ekaterina, "Data quality of platforms and panels for online behavioral research," *Behavior Research Methods*, pp. 1–20, 2021.

[70] B. D. Douglas, P. J. Ewell, and M. Brauer, "Data quality in online human-subjects research: Comparisons between mturk, prolific, cloudresearch, qualtrics, and sona," *Plos one*, vol. 18, no. 3, e0279720, 2023.

[71] J. Prokić, Ç. Çöltekin, and J. Nerbonne, "Detecting shibboleths," in *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, 2012, pp. 72–80.

[72] T. Javed *et al.*, *Svarah: Evaluating english asr systems on indian accents*, 2023. arXiv: 2305.15760 [cs.CL].

[73] C. M. Meyer and I. Gurevych, *Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography*. 2012.

[74] T. Ylonen, "Wiktextract: Wiktionary as machine-readable structured data," in *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC)*, European Language Resources Association (ELRA), 2022.

[75] A. Warstadt *et al.*, "Blimp: The benchmark of linguistic minimal pairs for english," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 377–392, 2020.

[76] T. Pham, T. Bui, L. Mai, and A. Nguyen, "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*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., Aug. 2021, pp. 1145–1160.

[77] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[78] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[79] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2017, pp. 43–1.

[80] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi, "Data decisions and theoretical implications when adversarially learning fair representations," *arXiv preprint arXiv:1707.00075*, 2017.

[81] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

[82] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, Jan. 2016.

[83] V. Mirjalili, S. Raschka, and A. Ross, "Privacynet: Semi-adversarial networks for multi-attribute face privacy," *IEEE Transactions on Image Processing*, vol. 29, pp. 9400–9412, 2020.

[84] J. Platt and A. Barr, "Constrained differential optimization," in *Neural Information Processing Systems*, 1987.

[85] A. Conneau *et al.*, "XNLI: Evaluating cross-lingual sentence representations," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 2475–2485.

[86] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[87] J. Pfeiffer, I. Vulić, I. Gurevych, and S. Ruder, "Mad-x: An adapter-based framework for multi-task cross-lingual transfer," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7654–7673.

[88] N. Houlsby *et al.*, "Parameter-efficient transfer learning for nlp," in *International conference on machine learning*, PMLR, 2019, pp. 2790–2799.

[89] T. B. Brown *et al.*, "Language models are few-shot learners," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20, Vancouver, BC, Canada: Curran Associates Inc., 2020, ISBN: 9781713829546.

[90] W. Held, D. Hall, P. Liang, and D. Yang, *Relative scaling laws for llms*, 2025. arXiv: 2510.24626 [cs.CL].

[91] E. Rolf, T. T. Worledge, B. Recht, and M. Jordan, "Representation matters: Assessing the importance of subgroup allocations in training data," in *International Conference on Machine Learning*, PMLR, 2021, pp. 9040–9051.

[92] J. Wei *et al.*, *Emergent Abilities of Large Language Models*, 2022. arXiv: 2206.07682 [cs.CL].

[93] Z. Du, A. Zeng, Y. Dong, and J. Tang, "Understanding Emergent Abilities of Language Models from the Loss Perspective," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[94] A. J. Van Deursen and J. A. Van Dijk, "The Digital Divide Shifts to Differences in Usage," *New media & society*, vol. 16, no. 3, pp. 507–526, 2014.

[95] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 610–623.

[96] M. Graham, B. Hogan, R. K. Straumann, and A. Medhat, "Uneven Geographies of User-generated Information: Patterns of Increasing Informational Poverty," *Annals of the Association of American Geographers*, vol. 104, no. 4, pp. 746–764, 2014.

[97] K. Naggita, J. LaChance, and A. Xiang, "Flickr Africa: Examining Geo-Diversity in Large-Scale, Human-Centric Visual Data," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '23, Montréal, QC, Canada: Association for Computing Machinery, 2023, pp. 520–530, ISBN: 9798400702310.

[98] Y. Elazar *et al.*, *What's In My Big Data?* 2023. arXiv: 2310.20707 [cs.CL].

[99] A. Yang *et al.*, *Qwen3 Technical Report*, 2025. arXiv: 2505.09388 [cs.CL].

[100] DeepSeek-AI *et al.*, *DeepSeek LLM: Scaling Open-Source Language Models with Longtermism*, 2024. arXiv: 2401.02954 [cs.AI].

[101] A. Grattafiori *et al.*, *The Llama 3 Herd of Models*, 2024. arXiv: 2407.21783 [cs.AI].

[102] N. Roberts, N. Chatterji, S. Narang, M. Lewis, and D. Hupkes, *Compute-Optimal Scaling of Skills: Knowledge vs Reasoning*, 2025. arXiv: 2503.10061 [cs.LG].

[103] T. Porian, M. Wortsman, J. Jitsev, L. Schmidt, and Y. Carmon, *Resolving Discrepancies in Compute-Optimal Scaling of Language Models*, 2025. arXiv: 2406.19146 [cs.LG].

[104] K. Wen, D. Hall, T. Ma, and P. Liang, *Fantastic Pretraining Optimizers and Where to Find Them*, 2025. arXiv: 2509.02046 [cs.LG].

[105] N. Kandpal *et al.*, *The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text*, 2025. arXiv: 2506.05209 [cs.CL].

[106] J. Li *et al.*, "DataComp-LM: In search of the next generation of training sets for language models," in *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[107] D. Su *et al.*, *Nemotron-CC: Transforming Common Crawl into a Refined Long-Horizon Pretraining Dataset*, 2025. arXiv: 2412.02595 [cs.CL].

[108] S. Greenbaum, *Comparing English Worldwide: The International Corpus of English*. Oxford University Press, Aug. 1996, ISBN: 9780198235828.

[109] J. W. Rae *et al.*, *Scaling Language Models: Methods, Analysis & Insights from Training Gopher*, 2022. arXiv: 2112.11446 [cs.CL].

[110] W. Held, B. Paranjape, P. S. Koura, M. Lewis, F. Zhang, and T. Mihaylov, *Optimizing pretraining data mixtures with llm-estimated utility*, 2025. arXiv: 2501.11747 [cs.CL].

[111] Y. Ju, F. Zhao, S. Chen, B. Zheng, X. Yang, and Y. Liu, "Technical report on conversational question answering," *ArXiv preprint*, vol. abs/1909.10772, 2019.

[112] I. Loshchilov and F. Hutter, *Decoupled Weight Decay Regularization*, 2019. arXiv: 1711.05101 [cs.LG].

[113] T. Xie *et al.*, "Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models," *ArXiv preprint*, vol. abs/2201.05966, 2022.

[114] G. Yang *et al.*, *Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer*, 2022. arXiv: 2203.03466 [cs.LG].

[115] Y. You *et al.*, *Large Batch Optimization for Deep Learning: Training BERT in 76 Minutes*, 2020. arXiv: 1904.00962 [cs.LG].

[116] S. Malladi, K. Lyu, A. Panigrahi, and S. Arora, *On the SDEs and Scaling Rules for Adaptive Gradient Algorithms*, 2024. arXiv: 2205.10287 [cs.LG].

[117] S. McCandlish, J. Kaplan, D. Amodei, and O. D. Team, *An Empirical Model of Large-Batch Training*, 2018. arXiv: 1812.06162 [cs.LG].

[118] M. Marek, S. Lotfi, A. Somasundaram, A. G. Wilson, and M. Goldblum, *Small Batch Size Training for Language Models: When Vanilla SGD Works, and Why Gradient Accumulation Is Wasteful*, 2025. arXiv: 2507.07101 [cs.LG].

[119] S. Hu *et al.*, "MiniCPM: Unveiling the Potential of Small Language Models with Scalable Training Strategies," in *First Conference on Language Modeling*, 2024.

[120] K. Wen, Z. Li, J. Wang, D. Hall, P. Liang, and T. Ma, *Understanding Warmup-Stable-Decay Learning Rates: A River Valley Loss Landscape Perspective*, 2024. arXiv: 2410.05192 [cs.LG].

[121] A. Defazio, *Why Gradients Rapidly Increase Near the End of Training*, 2025. arXiv: 2506.02285 [cs.LG].

[122] K. Liang, L. Chen, B. Liu, and Q. Liu, *Cautious Optimizers: Improving Training with One Line of Code*, 2025. arXiv: 2411.16085 [cs.LG].

CHAPTER 6

APPENDICES

6.1 Multi-VALUE Models & Hyperparameters

CoQA We use the base versions of BERT [52] and RoBERTa [53] on dialect variants of the CoQA task, following the Rationale Tagging Multi-Task setup of [111] to adapt these models to the CoQA setup which includes *Yes, No,* and *Unknown* responses in addition to extractive answers. Each model was trained on an Nvidia GeForce RTX 2080 Ti for approximately 6 hours. For each model and dialect, we fine-tune using AdamW [112] for 2 epochs with a batch size of 16 and a learning rate $3e-5$.

Semantic Parsing. Following [113], for T5-base we adopted the AdamW optimizer, while Adafactor was used for T5-3B and the two BART models. We used NVIDIA A100 to train these models with T5-3b, BART-large, T5-base, and BART-base using 8 GPUs for 52 hours, 4 GPUs for 32 hours, 4 GPUs for 4 hours, 4 GPU for 13 hours respectively. We set the learning rate at 5e-5 for T5 models and 1e-5 for BARTs. We fixed the batch size at 32 when fine-tuning T5-BASE and BARTs. As for the extremely large T5-3B, we configured a batch size of 64 to speed up convergence and utilised DeepSpeed to save memory. Linear learning rate decay was used for all models.

Machine Translation. We evaluate the NLLB Translation Model at two distilled scales: 615M and 1.3B [57]. Evaluation was done on an Nvidia GeForce RTX 2080 Ti and takes less than 10 minutes. The NLLB model is designed for many-to-many translation with low-resource language communities and is trained on a large corpus mined from the internet, rather than exclusively human aligned translations. We choose this model to give us an estimate of the performance of large scale translation products available to users.

6.2 Intrinsic Dialect Benchmark Construction

6.2.1 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[1].

We use a Wiktextract [74], 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 loanwords and calques, and assess Challenge #2 by intersecting this list with Wiktionary's list of English Borrowed Terms[2]. 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.

6.2.2 Evaluating Modeling of SAsE Syntax

While existing work has evaluated the functional effects of Indian English syntax on downstream tasks [17], these assess the robustness of a model in the face of syntactic variation.

[1] List of Terms associated with SAsE on Wiktionary
[2] List of All English Borrowed Terms

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) [75], 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[3] aligned syntax and syntax aligned with SAmE or British English [65]. 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 [45], 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 [17] 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.

[3]Some of the feature of [65] 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

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.

6.3 IsoFLOP Hyperparameter Scaling

6.3.1 Architecture

Width. The hidden size d is restricted to multiples of 128, reflecting accelerator block sizes. d ranges from 512 to 4096 in increments of 128 for small budgets (up to 9×10^{18} FLOPs) and increments of 256 for larger budgets.

Depth. Depth is determined by a log-corrected rule dependent on width:

$$L = \frac{d}{\kappa + \theta \log_2 d}.$$

The parameters θ and κ are adjusted to align depth-to-width ratios with those reported in Hoffmann *et al.* [31], which require empirical alignment to set the appropriate parameter values.

Attention heads and MLP Ratio. Attention head size and MLP ratio follow standard practice [86]. We set $n_{\text{heads}} = d/128$, so each head spans 128 dimensions, and use conventional multi-headed attention. The feed-forward dimension is fixed at $4d$, as in most open models.

6.3.2 Optimization

Batch size and steps. To maintain comparability across runs, we target a training length of 2^{16} steps [114]. For a token budget T, the batch size B is computed via $T = B \cdot 2^{16}$ and rounded to the nearest power of two for efficiency. The step count is then adjusted to

recover T.

Learning rate. Given batch size B and hidden size d, the learning rate is defined as

$$\eta = \eta_{base} \frac{\sqrt{B}}{d}.$$

This scaling, consistent with μP analysis [114] and large-batch rules [115, 116], decreases with width and increases with batch size. In practice, $\eta \geq 0.01$ causes reproducible loss spikes, consistent with McCandlish *et al.* [117]. Runs with such learning rates are forced to use smaller batch sizes until $\eta \leq 0.01$, which extends training length and alters dependent hyperparameters. These longer runs are likely sub-optimally tuned, but this mostly affects small models trained at large token budgets, which we do not expect to be compute optimal regardless.

Miscellaneous. We set $\beta_2 = 0.95$, with smaller batch sizes using reduced decay according to Marek *et al.* [118]. Other settings are fixed: $\beta_1 = 0.95$, $\epsilon = 10^{-15}$, weight decay $= 0.1$, gradient clipping at norm 1.0. A Warmup–Stable–Decay schedule is used [119, 120], with 5% warmup and 20% linear decay.

Stability modifications to AdamW. Training uses AdamW [112] augmented with AdamC [121] and Caution [122]. AdamC corrects weight-decay/normalization interactions that otherwise increase gradient norms late in training, and Caution suppresses momentum updates conflicting with gradient direction. These interventions improve smoothness, but their necessity indicates that stability is not inherent to the base configuration. .tex