THESIS PROPOSAL:
IMPROVING THE ROBUSTNESS OF NATURAL LANGUAGE PROCESSING TO DIALECTS AND LANGUAGE VARIANTS

Thesis Proposal
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

March 2025

THESIS PROPOSAL:
IMPROVING THE ROBUSTNESS OF NATURAL LANGUAGE PROCESSING TO DIALECTS AND LANGUAGE VARIANTS

Thesis Proposal committee:

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

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

Dr. Judy Hoffman
School of Interactive Computing
Georgia Institute of Technology

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

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, Completed Work)

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 evaluations 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, Completed Work)

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, Completed & Proposed Work)

Concurrent to my work on methods to adapt PLMs to dialects, a significant amount of research has focused on the predictable improvements that are derived by increasing the *scale* of language model pretraining. This research direction is often grounded in the concept of *scaling laws* [30, 31] — empirically fit functions of model performance with respect to the scale of model size and training data of the underlying model. As such, a natural question for my work on *adapting* language models for dialect robustness is ”Are these still necessary with larger scale models?”

Neither my own work, nor any other work in the field, currently addresses this question. Theoretical work on scaling laws for ”minority data” [32] is inconclusive on this question, as theoretical findings show that scale can cause disparities between groups to diminish, hold constant, or even increase depending on the interaction effects. I plan to provide empirical results to answer conclusively whether scale can decrease dialect disparity without other interventions. In this proposal, I will describe my plan to achieve this, including the ongoing collaborative work I have with the Stanford Center for Research on Foundation Models to produce an open-source set of checkpoints to fit scaling laws effectively.

Finally, I will explore how much more efficiently dialect disparity can be diminished using the methods I have developed throughout my PhD rather than simple scaling. This will cover synthetic data augmentation such as Multi-VALUE, post-training methods such as TADA, and data curation methods that I have developed for multi-domain robustness [33].

1.5 Thesis Statement

English as a global language—spoken by billions across continents—is rich with variation. Despite the number of speakers of other variants and dialects, most language technologies primarily serve Standard American English speakers, creating systematic barriers for other dialect communities. My research establishes empirical evidence for these disparities through novel controlled experiments and user experience studies spanning multiple English varieties. Building on these findings, I develop computationally efficient adaptation techniques that enhance dialect robustness without requiring dialect-specific annotations for each downstream task. To complete my thesis, I intend to study the relationship between model scale and dialect performance, demonstrating that intentional dialectal interventions offer more efficient pathways to dialect robustness than scaling alone. These contributions advance both the theoretical understanding of language variation as a dimension of NLP fairness and provide practical methods for building language technologies that serve English in all its forms.

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 [34, 35, 36, 37, 38], NLI [18], dependency parsing, POS tagging [20, 22], and downstream applications [39]. 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.

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 [40] — 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` [41] and `stanza` [42] 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 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

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.

speakers of a particular dialect. This is a key differentiating feature from work using unvalidated synthetic features without clear correspondence to real world variation [43].

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 [44, 45].

For brevity in this proposal, I focus on our results for conversation question answering based on CoQA from Reddy *et al.* [46]. 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[47] and RoBERTa[48] 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[49] and BART[50], for semantic parsing on the SPIDER benchmark [51] and models specified for machine translation [52] on the WMT-19 benchmark [53]. 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.

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 [54], 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 [55].

Despite South Asia having an enormous English speaking community [56, 57] and extensive NLP research [58, 59, 60, 61, 62, 63, 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 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 [64, 65], balanced recruitment, and screening capabilities. This enabled us to

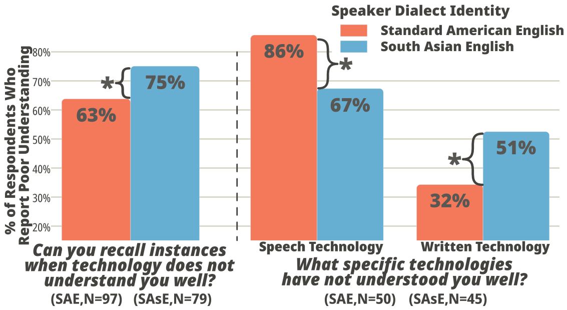

Figure 2.1: **Survey responses** to the questions *"Can you recall instances when technology does not understand you well?"* and *"What specific technologies have not understood you well?"*. * denotes significance at $P < 0.05$ using a Barnard Exact test.

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.

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 [66] 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 2.1) 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 Assistants or Automated Customer Service) or primarily text-based (such as Chatbots or Search Engines). SAsE speakers are significantly (+19%, P=0.026) more likely than their SAmE counterparts to list at least one written technology like ChatGPT, search engines, and Grammarly and significantly (-19%, P=0.012) less likely to list at least one spoken technology such as Siri, Alexa, and automated phone services. This finding indicates that the empirical disparities noted in prior works 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 [67]. 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.

As an intrinsic assessment of lexical understanding, we scraped 724 stand-alone terms and 317 loanwords from other South Asian languages from Wiktionary [68, 69] 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.* [70] with 110 sentences aligned between SAE and Indian English [60].

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 2.2), 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. No-

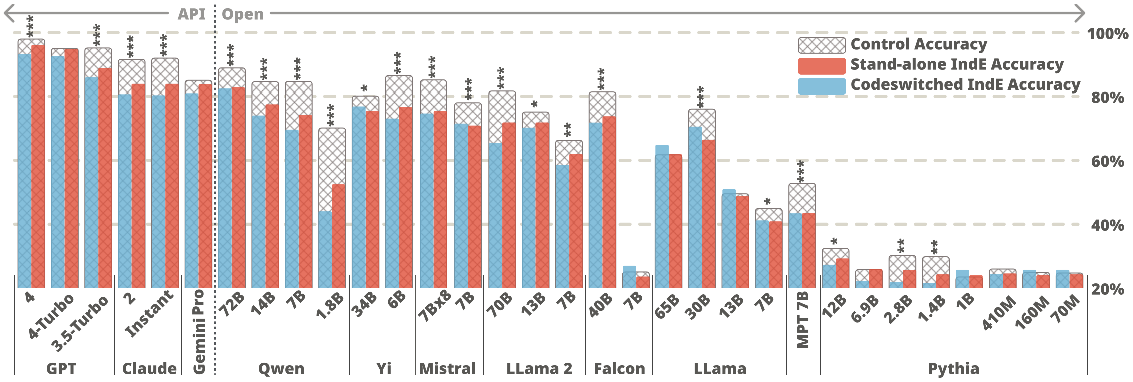

Figure 2.2: **Results for Wiktionary Benchmarks of both SAsE and Unmarked Lexical Knowledge**. *, **, and *** denote cases where overall performance is worse at P¡0.05, P¡0.01, and $P < 0.001$ respectively by a Bootstrap test. Control accuracy is for terms without any regional affiliation on Wiktionary.

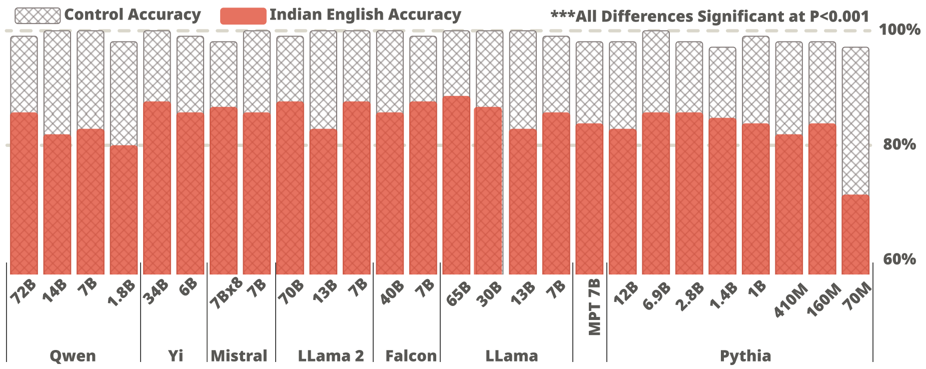

Figure 2.3: **Results for Minimal Pair Benchmark of both Indian and SAmE Syntactic Knowledge**. While the smallest models consistently perform nearly perfectly on the SAmE control, even the largest models perform significantly ($P < 0.001$) worse on the Indian English evaluation. Significance computed using a Bootstrap significance test.

tably, Indian English performance correlates strongly with control performance (p=0.98).

Regarding syntactic processing (shown in Figure 2.3), all evaluated models demonstrate near-perfect performance on SAE syntax while exhibiting statistically significant 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 [71].

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) [72], equality of opportunity (the false positive rate should be independent of Z) [73] , and equalized odds(that odds of misclassifications should be equal across groups) [74] 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-

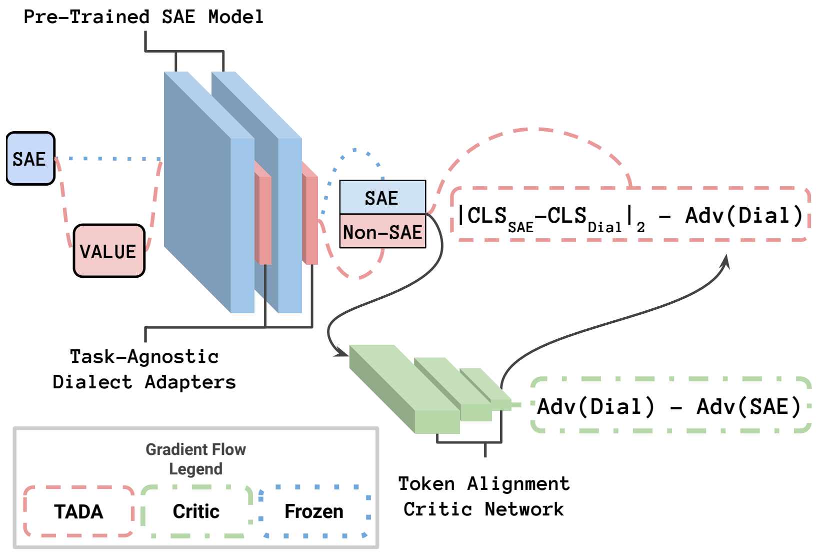

Figure 3.1: **Task-Agnostic Adapter training flow** with both sequence and token level alignment loss between SAE and a target dialect. When stacked before task-specific SAE adapters, TADA provides dialect robustness for the target task.

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 using adversarial methods [75, 76]. Beyond fairness, using adversarial learning to remove undesirable features had been discovered and applied separately in transfer learning [77] and privacy preservation [78]. As shown in Figure 3.1, 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 [76] 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 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 [79]. 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 [80]. 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 information 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:

$$L_{seq} = |\vec{CLS}_{sae} - \vec{CLS}_{dial}|_2 \tag{3.1}$$

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 [81]. 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 [60] 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 [82] to minimize the combined loss $L_{TADA} = L_{seq} + L_{tok}$. The full training procedure is shown in Figure 3.1.

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 [83]. 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 benchmark. However, TADA performs 4% worse on average than task-specific VALUE-

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 [48]. 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 3.2: **Multi-Dialectal** evaluation results (Mean across all tasks) for 4 Non-SAE Dialect Variants of GLUE created using Multi-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 3.2. 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 AND IMPROVING DIALECT ROBUSTNESS DURING PRETRAINING

While the prior section focuses on creating adaptation modules for existing models that are plug-and-play, these methods require each individual practitioner to seek out dialect robustness interventions. Unfortunately, most practitioners are unlikely to consider dialect robustness when building real-world applications, especially as PLMs are increasingly adopted as a general tool in real-world software where aspects such as robustness are unlikely to be systematically tested [24].

As this limits the practical impacts of my work, it would be extremely valuable to understand how dialect robustness can be embedded in PLMs by design, beginning in the pretraining stage. While the methods I have worked on can be adapted to achieve this, a natural question from a skeptic is: Will dialect disparities naturally diminish with increased scale, making these methods irrelevant? Without a convincing answer to this question, developers are unlikely to prioritize investments into dialect robustness specifically alongside where scale has been shown to not yield improvements, such as toxicity [84]. Therefore, as the final stage of my thesis, I propose testing scaling hypotheses for dialects, leveraging the experimental framework of scaling laws to quantify the value of my previous contributions in terms of compute efficiency and create a framework for forecasting future robustness to guide research priorities in dialectal NLP.

4.1 Scaling Laws and Dialect Performance

Scaling laws, as introduced by Kaplan *et al.* [30], forecast how model performance improves as a power-law function of increased resources—whether measured in parameters, training tokens, or compute. These models assume that scaling behavior follows the form

$L(X) \approx (X_0/X)^{\alpha}$, where L is the final loss as a function of the aspect of scale being studied X, where X_0 and α determine the rate of improvement.

In Hoffmann *et al.* [31], the functional form was extended to understand how the amount of training data and model size interact when training a model according to a particular compute budget. This models scaling as a linear combination of scaling laws for both model (M) and training data (D) size, plus a term for the inherent entropy of the task:

$$L(M, D) \approx A + \frac{M_0}{M^{\beta}} + \frac{D_0}{D^{\alpha}} \tag{4.1}$$

By training models at a variety of M and D, finding the scaling ”law” is simply a matter of training a regression on the results of different model configurations. The advantage is that the resulting regression can be used forecast the performance of models of larger scales based on small scale runs. Similarly, given a particular loss, the regression model can be used to identify how much additional training data, model parameters, or overall FLOPs would be needed to achieve that performance.

This experimental procedure has allowed pretraining research to run more principled small-scale experiments for transfer to the large-scale models which dominate in applications. In my own work, I have utilized scaling projections to better compare data curation methods and their impacts on multiple downstream tasks [33].

However, in their current empirical form, scaling laws do not offer insights into whether dialect disparities diminish with scale as both major works on this front test scaling only on in-distribution data. Theoretical work by Rolf *et al.* [32] hypothesizes that minority data distributions splits the data scaling term $\frac{D_0}{D^{\alpha}}$ into two terms $\frac{D_{i,0}}{D_i^{\alpha_i}} + \frac{D_{t,0}}{D_t^{\alpha_t}}$ where t indicates terms corresponding to the shared loss across all data distributions and i indicates terms corresponding to the unique loss for a particular distribution.

The implications of this equation for our core question are unclear without empirically fir parameters. If $\alpha_t > \alpha_i$, the dialect disparity will decrease with scale while the opposite is true if the inequality is reversed. To answer the question about the relationship between scale and dialectal robustness, I have begun work to empirically fit dialectal scaling laws.

4.2 Proposal & Timeline

To address this gap, my final dissertation chapter will perfon the first comprehensive analysis of dialect performance as a function of pretraining scale. Working in collaboration with the Stanford Center for Research on Foundation Models (CRFM), I have already begun training models from scratch ranging from 200M parameters to 1.4B parameters on commonly used pretraining data [85].

One core challenge of performing scaling law research in academia historically has been compute requirements needed to train a large enough sample size of even small models. By working with CRFM, I have been able to utilize large amounts of compute which would be otherwise inaccessible to me. Furthermore, along with my CRFM collaborators, we have shown in early experiments that accurate scaling laws can be fit using stable decay learning rate schedules [86] which allows computation to be shared between models of the same size trained for different lengths, while Hoffmann *et al.* [31] shows that the same is not possible for models trained with cosine learning rates.

Once this training runs complete (est. May 2025), I intend to evaluate their scaling behaviors empirically on raw dialectal language modeling [87], dialectal benchmarks [19, 88], and measures of bias triggered by dialect [89]. First, I will establish the impacts of scaling alone in these dimensions to determine whether scale indeed decreases dialectal disparities. Then, I intend to compare the benefits of scale to the benefits of my prior dialectal methods [27, 17] and dialectal data curation based on my domain adaptation work [33]. These findings will allow me to better communicate the outcomes of my entire research trajectory, not just with respect to the field at the time of its completion, but in a language that enables projecting these benefits into the future.

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] 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.

[33] 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]`.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

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

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

[42] 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]`.

[43] 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*.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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

[49] 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.

[50] 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.

[51] 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.

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

[53] 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.

[54] 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.

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

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

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

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

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

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

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

[69] 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.

[70] 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.

[71] 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.

[72] 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.

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

[74] 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.

[75] 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.

[76] 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.

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

[78] 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.

[79] J. Platt and A. Barr, “Constrained differential optimization,” in *Neural Information Processing Systems*, 1987.

[80] 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.

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

[82] 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.

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

[84] A. Birhane, V. U. Prabhu, S. Han, and V. N. Boddeti, “On hate scaling laws for data-swamps,” *CoRR*, 2023.

[85] L. Soldaini *et al.*, “Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 15 725–15 788.

[86] 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,” *arXiv preprint arXiv:2410.05192*, 2024.

[87] S. Greenbaum, “Ice: The international corpus of english,” *English Today*, vol. 7, no. 4, pp. 3–7, 1991.

[88] 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.

[89] V. Hofmann, P. R. Kalluri, D. Jurafsky, and S. King, “Ai generates covertly racist decisions about people based on their dialect,” *Nature*, vol. 633, no. 8028, pp. 147–154, 2024.