Stanford | NLP

Dynamic Aggregation and Auto-Discovery of Linguistic Features

Yanchen Liu

Stanford | NLP

DADA: Dialect Adaptation via Dynamic Aggregation of Linguistic Rules

Yanchen Liu, William Held, Diyi Yang

will appear on EMNLP 2023

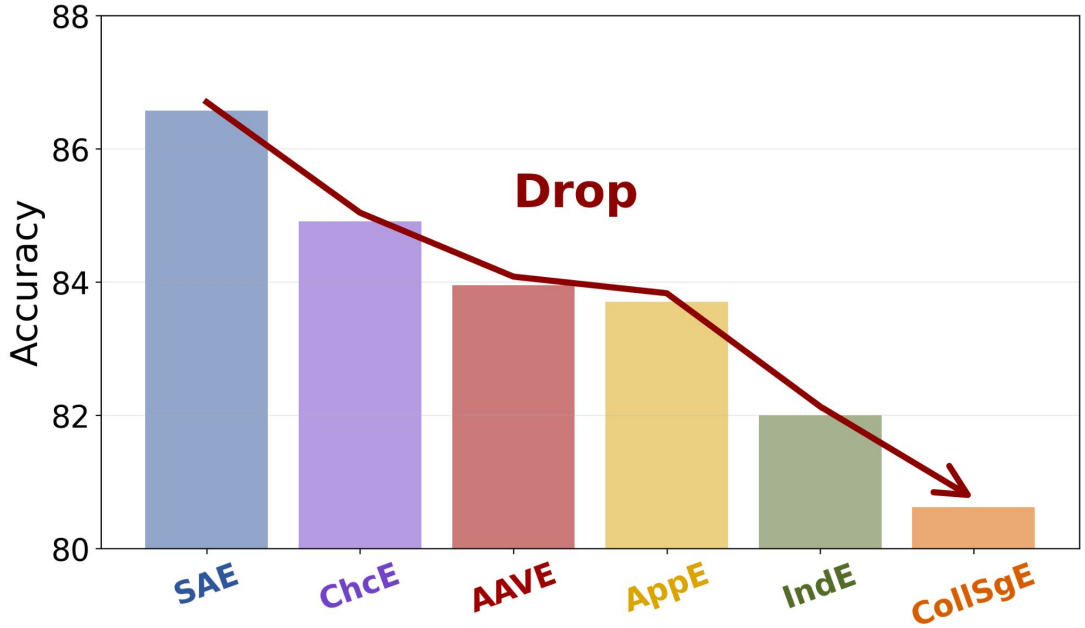
Non-Standard Linguistic Features

- language usage that deviates from the conventions
- often associated with specific social or cultural groups

Dialect: A Group of Non-Standard Linguistic Variations in a Language

Fails on Dialects

- Existing models mainly focus on Standard American English (**SAE**)
- Significant **performance drop**↓when applied to English Dialects



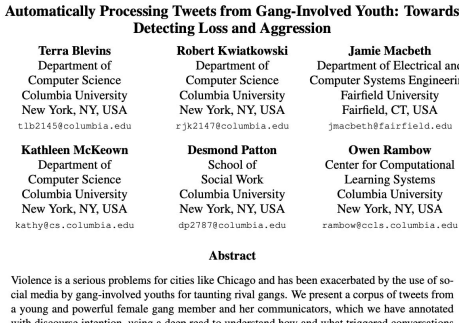
Previous Work on Dialect Adaptation

Stanford | NLP

Mainly focused on targeted adaptation to **a specific dialect**

- Human Annotation (Blevins et al., 2016, Blodgett et al., 2018)
- Weak Supervision (Jorgensen et al. 2016, Jurgens et al. 2017)
- Alignment (TADA 2023)

⚠ **Highly accurate dialect identification systems are required!**



Automatically Processing Tweets from Gang-Involved Youth: Towards Detecting Loss and Aggression

Terra Blevins
Department of Computer Science
Columbia University
New York, NY, USA
tlb2145@columbia.edu

Robert Kwiatkowski
Department of Computer Science
Columbia University
New York, NY, USA
rjk2147@columbia.edu

Jamie Macbeth
Department of Electrical and Computer Systems Engineering
Fairfield University
Fairfield, CT, USA
jmacbeth@fairfield.edu

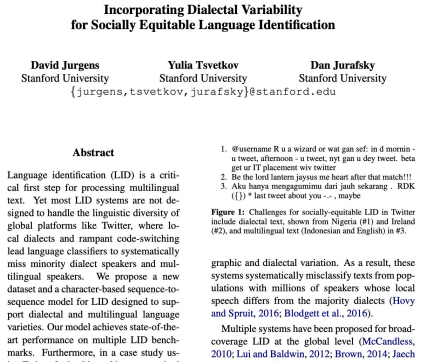
Kathleen McKeown
Department of Computer Science
Columbia University
New York, NY, USA
kathy@cs.columbia.edu

Desmond Patton
School of Social Work
Columbia University
New York, NY, USA
dp2787@columbia.edu

Owen Rambow
Center for Computational Learning Systems
Columbia University
New York, NY, USA
rambow@ccls.columbia.edu

Abstract

Violence is a serious problems for cities like Chicago and has been exacerbated by the use of social media by gang-involved youths for taunting rival gangs. We present a corpus of tweets from a young and powerful female gang member and her communicators, which we have annotated



Incorporating Dialectal Variability for Socially Equitable Language Identification

David Jurgens Stanford University
Yulia Tsvetkov Stanford University
Dan Jurafsky Stanford University
{jurgens,tsvetkov,jurafsky}@stanford.edu

Abstract

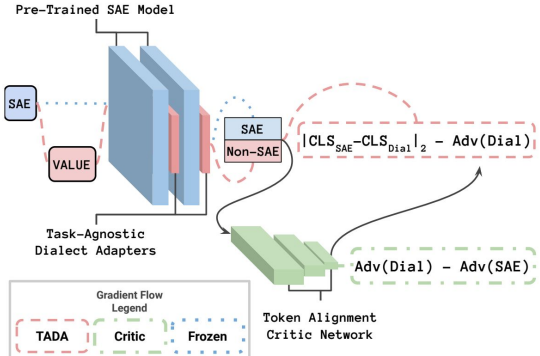
Language identification (LID) is a critical first step for processing multilingual text. Yet most LID systems are not designed to handle the linguistic diversity of global platforms like Twitter, where local dialects and rampant code-switching lead language classifiers to systematically miss minority dialect speakers and multilingual speakers. We propose a new dataset and a character-based sequence-to-sequence model for LID designed to support dialectal and multilingual language varieties. Our model achieves state-of-the-art performance on multiple LID benchmarks. Furthermore, in a case study us-

1. @username R u a wizard or wat gan sef: in d mornin - u tweet, afternoon - u tweet, nyt gan u dey tweet. beta get ur IT placement wiv twitter
2. Be the lord lantern jaysus me heart after that match!!!
3. Aku hanya mengagumimu dari jauh sekarang . RDK ({}) * last tweet about you -_- , maybe

Figure 1: Challenges for socially-equitable LID in Twitter include dialectal text, shown from Nigeria (#1) and Ireland (#2), and multilingual text (Indonesian and English) in #3.

graphic and dialectal variation. As a result, these systems systematically misclassify texts from populations with millions of speakers whose local speech differs from the majority dialects (Hovy and Spruit, 2016; Blodgett et al., 2016).

Multiple systems have been proposed for broad-coverage LID at the global level (McCandless, 2010; Lui and Baldwin, 2012; Brown, 2014; Jaech



However!!! Inherent **Flexibility** of Dialects

Stanford | NLP

- Flexible Boundaries

  => no highly accurate dialect identification systems available

- Vary Depending on Personal and Social Contexts

  => dialects do not neatly fit into predefined categories

🌟 **Accommodate the diversity of dialects from a Fine-Grained perspective Linguistic Features**

Method

Dialect Adaptation via Dynamic Aggregation

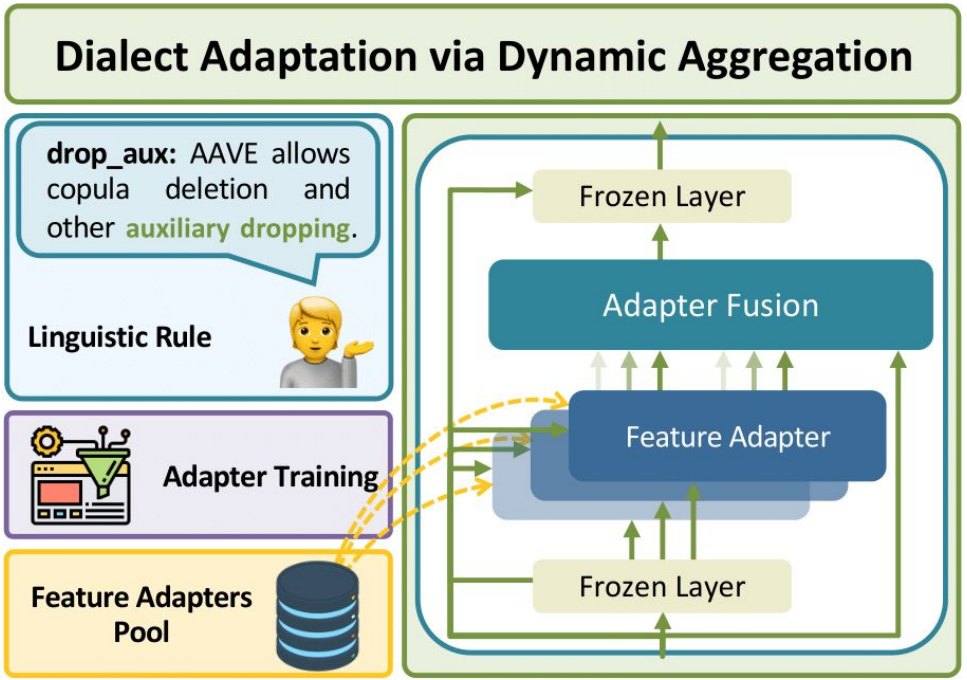
Stanford NLP

🏅 Modular and Dynamic

- ❏ Multi-Dialectal Robustness
  - ❏ Input Dialect-Agnostic
  - ❏ Personal- and Social-Contextual

🌟 Within Only 3 Steps

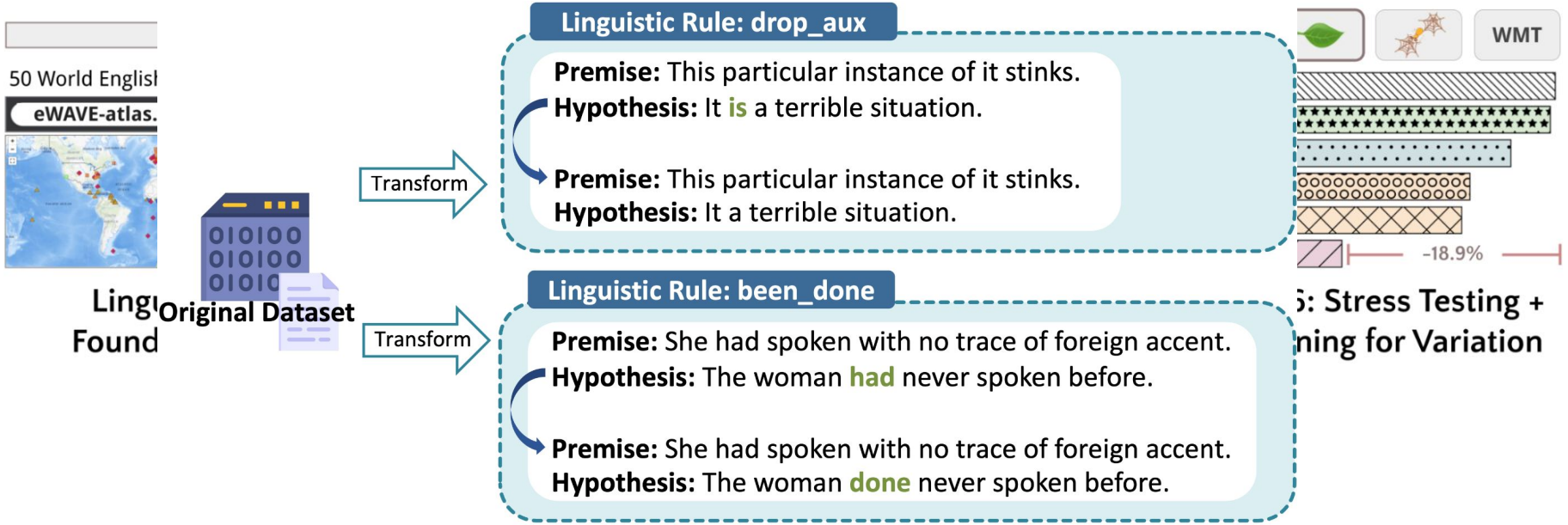
1. Synthetic Datasets Construction
2. Feature Adapter Training
3. Dynamic Aggregation



Step 1: Synthetic Datasets Construction

Stanford | NLP

Construct a transformed dataset for each non-standard (morphosyntactic) linguistic feature.

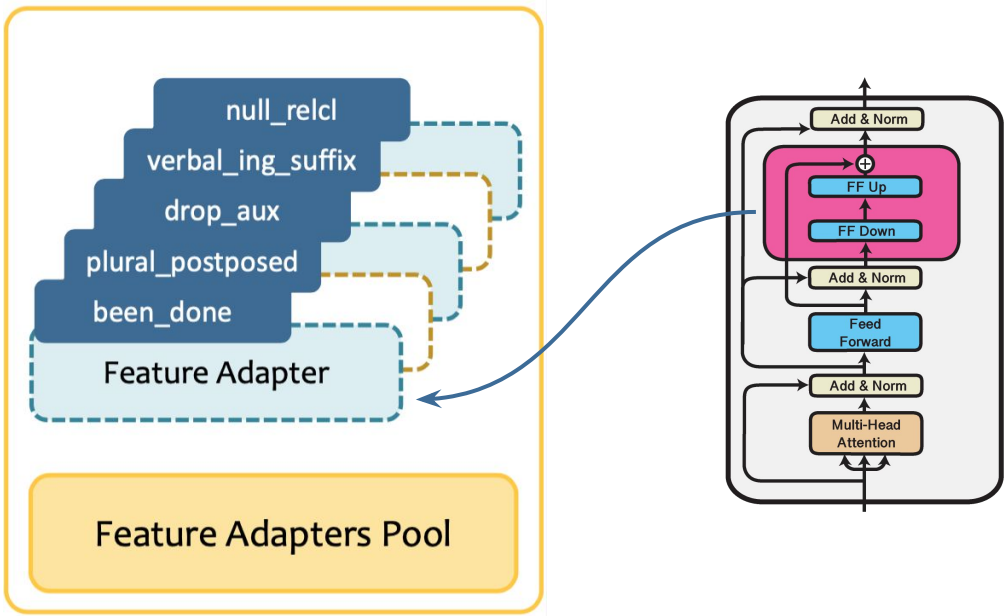


[2] Ziems et al. 2022. Multi-VALUE: Evaluating Cross-dialectal NLP. ACL 2023.

Step 2: Feature Adapter Training

Stanford | NLP

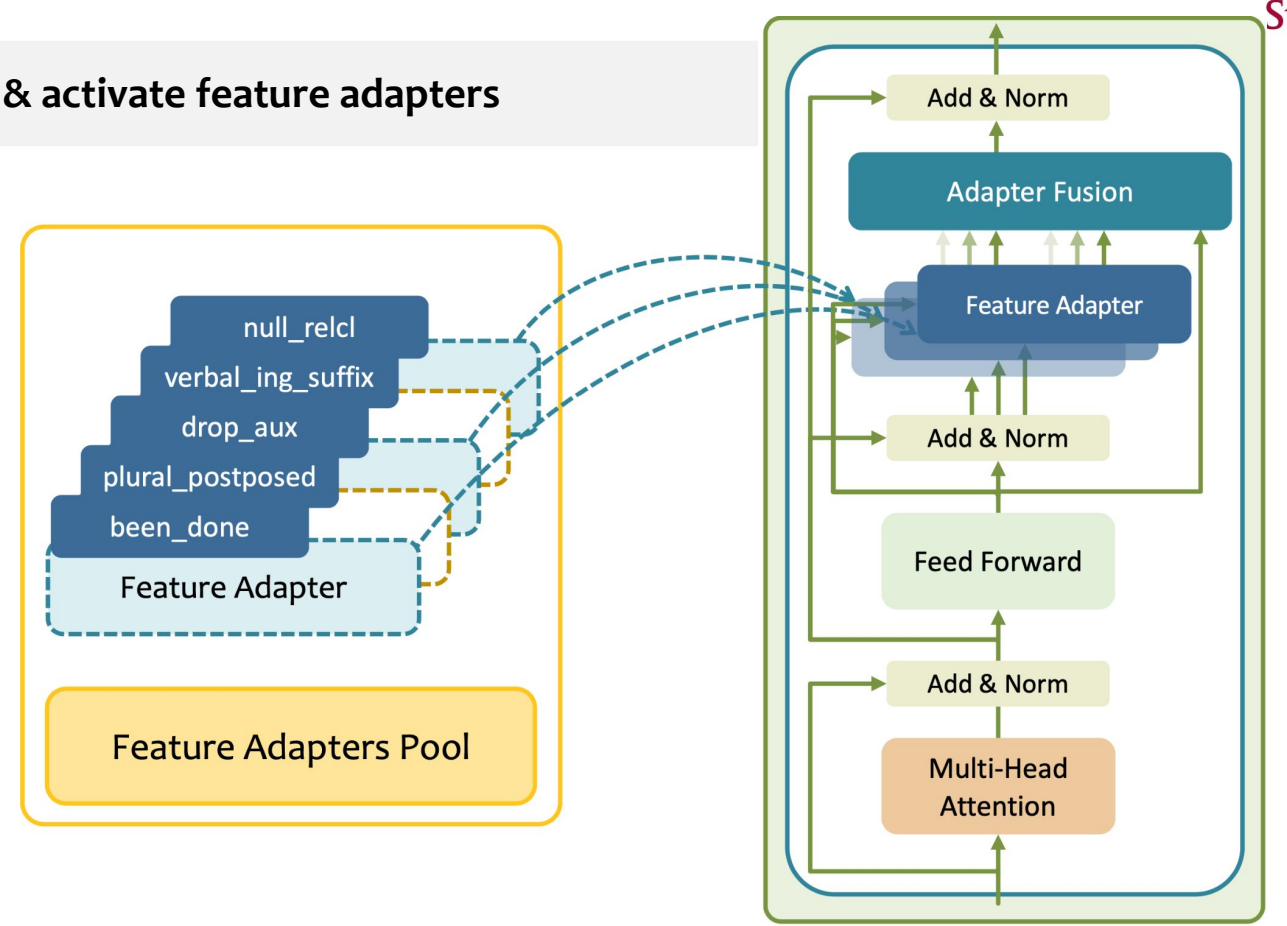
Train a feature adapter for each non-standard linguistic feature.



Step 3: Dynamic Aggregation

Stanford | NLP

🔖 **Aggregate & activate feature adapters dynamically.**

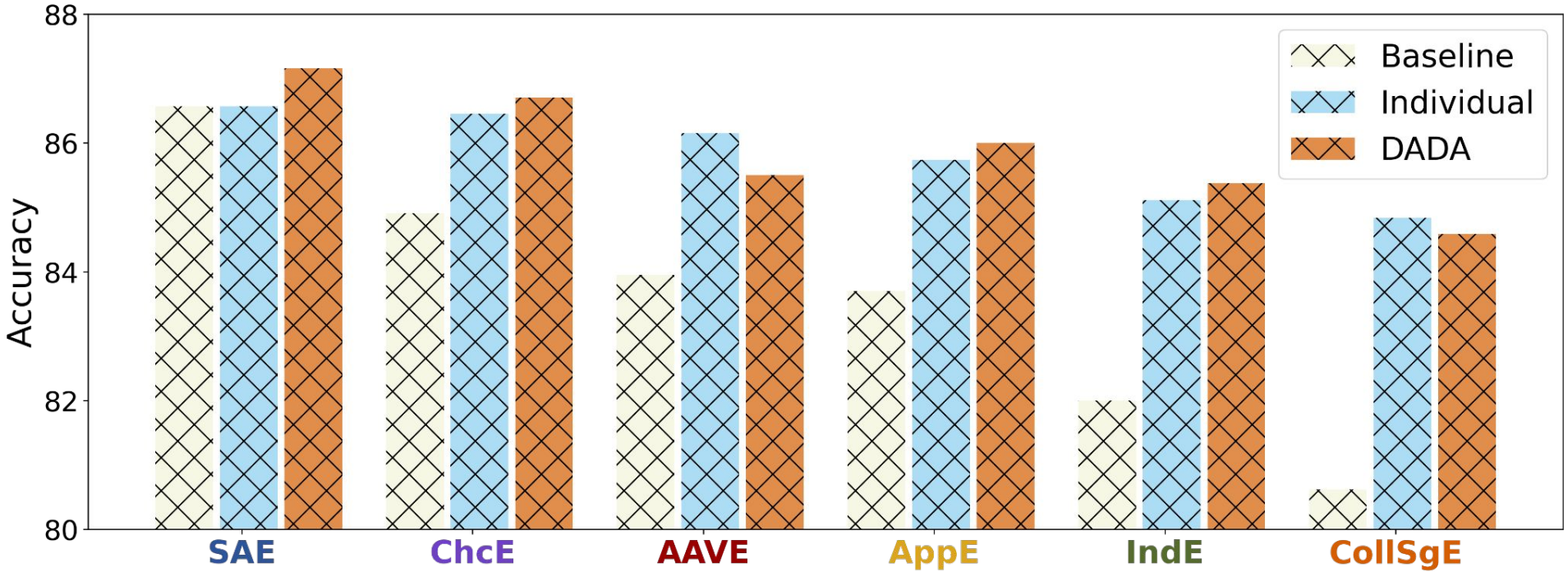


Experiments

1/ DADA Can Improve Multi-Dialectal Robustness

Stanford | NLP

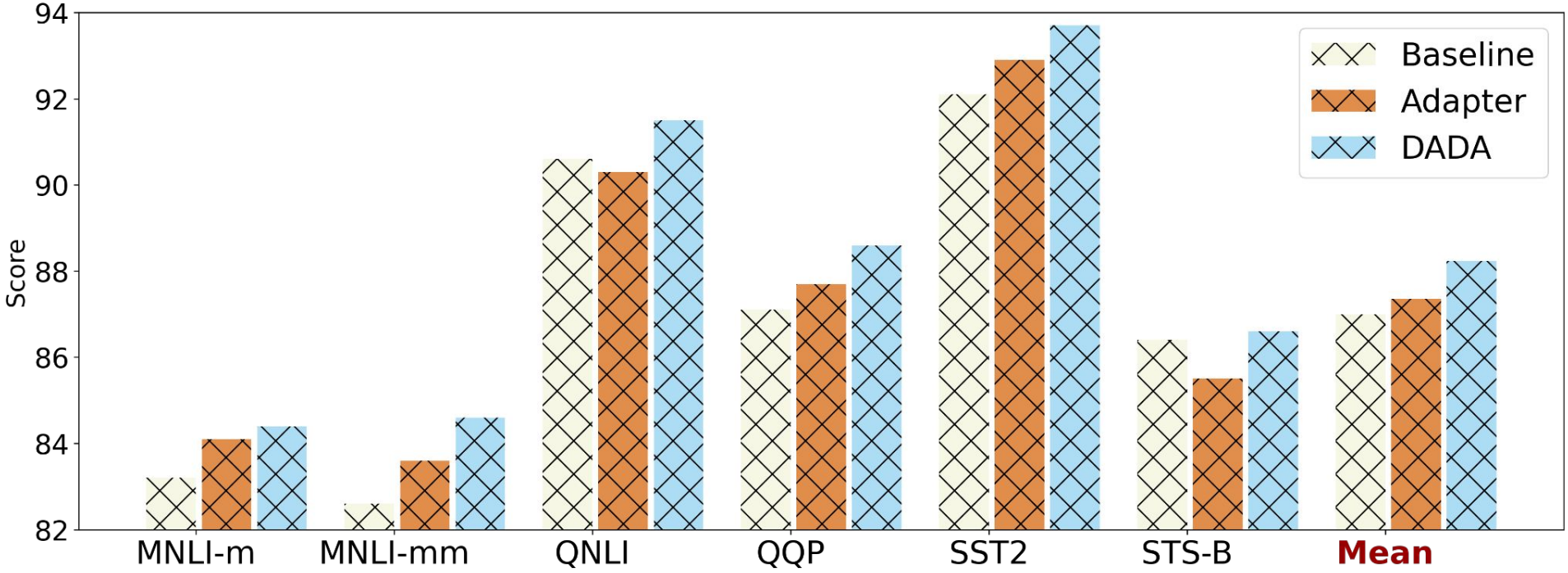
Adapt **SAE** model to multiple dialect variants simultaneously: **AppE**, **ChcE**, **CollSgE**, **IndE**, **AAVE**



2/ DADA Can Be Task-Agnostic

Stanford | NLP

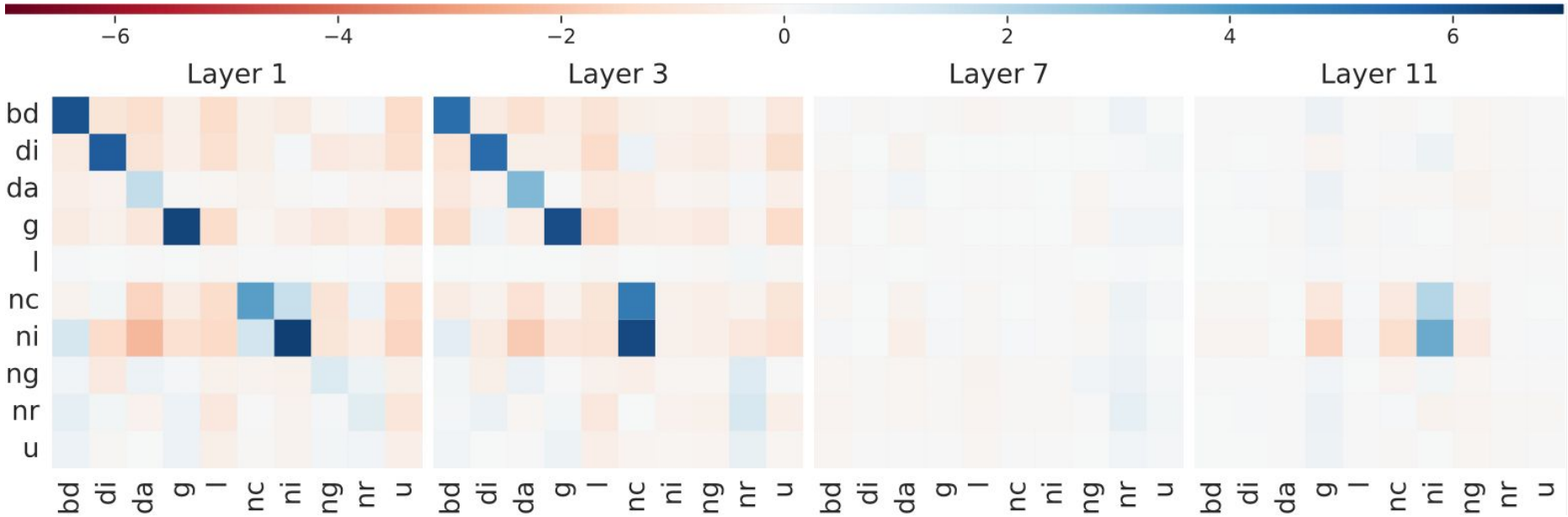
Adapt instruction-tuned **SAE** model to the dialect variants for multiple tasks



3/ DADA Has Great Interpretability!

Stanford | NLP

Correlation Coefficients for **AAVE** Adaptation



We use abbreviations for certain terms, such as "nc" for "negative concord."

Conclusion and Future Work

Dynamic Aggregation of Linguistic Features

Stanford | NLP

- A **Fine-grained** and **Modular** Method for Dialect Adaptation
  - Improve **Multi-Dialect** and **Multi-Task** Robustness
    - No need for highly accurate dialect identification systems
    - Taking personal and social context into account
    - Applicable to task-agnostic instruction-tuned LLMs
  - **Interpretability**, reusability and extensibility

But!!!

Stanford | NLP

Non-Standard Linguistic Features

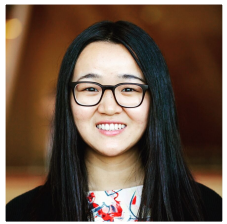
- ❏ are curated by linguists (eWAVE) and
- ❏ play a crucial role in a wide range of applications.

However, the manual curation of linguistic rules can be **expensive** and **expertise-intensive**.

Empower Linguistic Research with LLMs

Large Language Models Can Discover Linguistic Features

(ongoing)



Introduction

- Leverage LLMs to help:
  a. **Verify and Identify** non-standard linguistic features
     - given a corpus, identify the non-standard linguistic features
     - current scope:
       - 235 non-standard (morphosyntactic) English linguistic features from eWAVE
     - more in the future:
       - lexical, stylistic, etc.
       - more languages
  b. **Discover** new linguistic features/usages
     - or new usages
       - *e.g. I am shopping (at) Walmart*

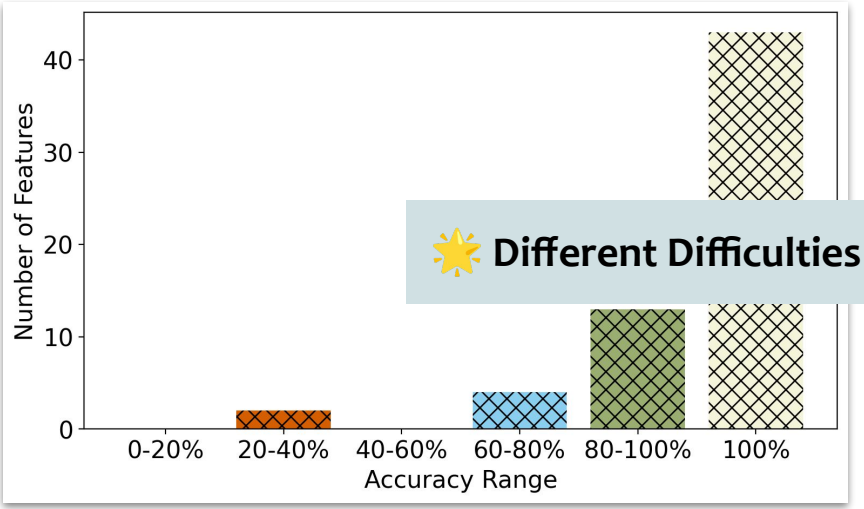
🌟 **Linguistic Feature Discovery**

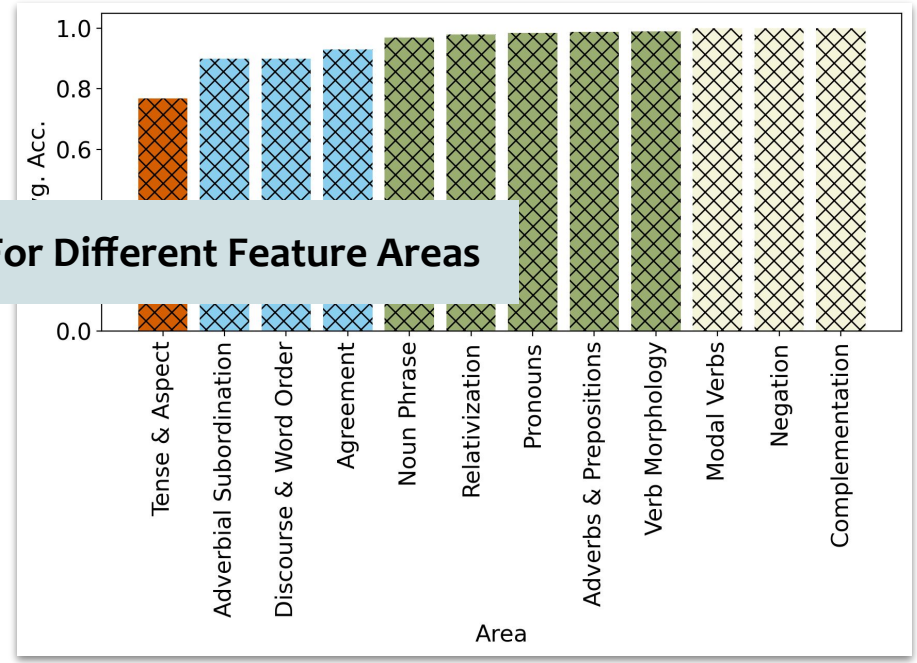
He won't do no harm. → ***Negative Concord***

1/ Multiple Choice

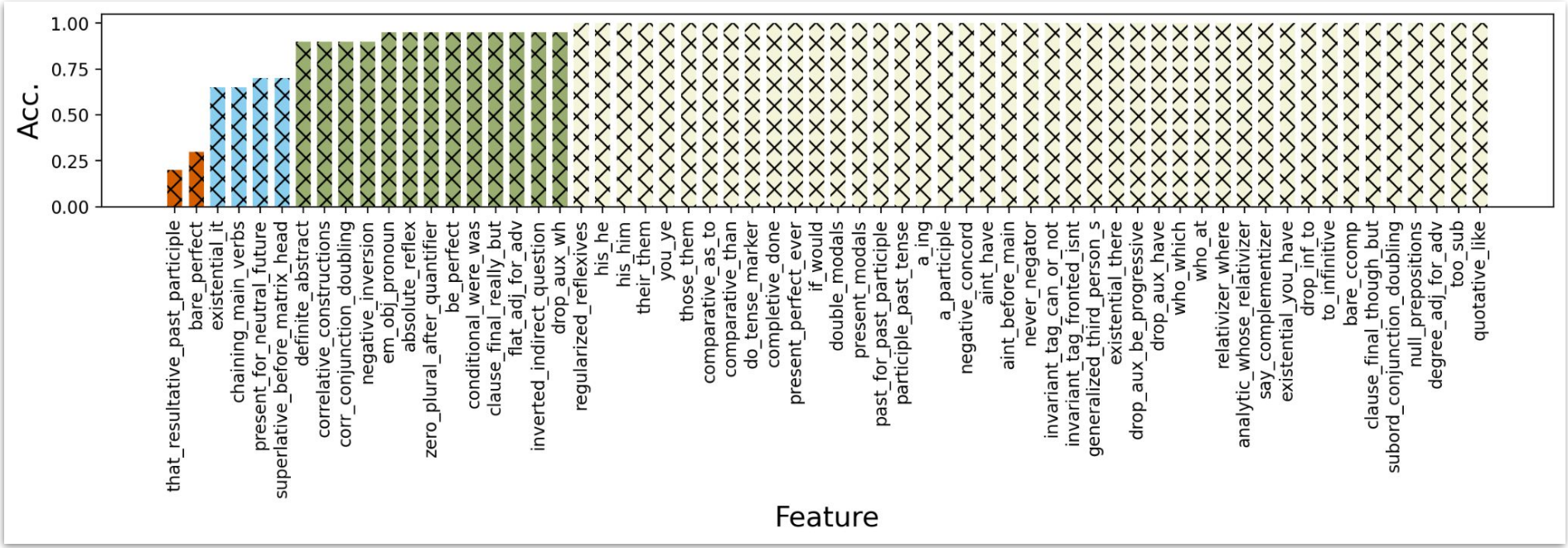
Setup1: ask LLM to identify one **single** linguistic feature at a time

- provide LLM a (synthetic) paragraph with a **single** linguistic feature, along with 5 choices





🌟 **Different Difficulties For Different Feature Areas**

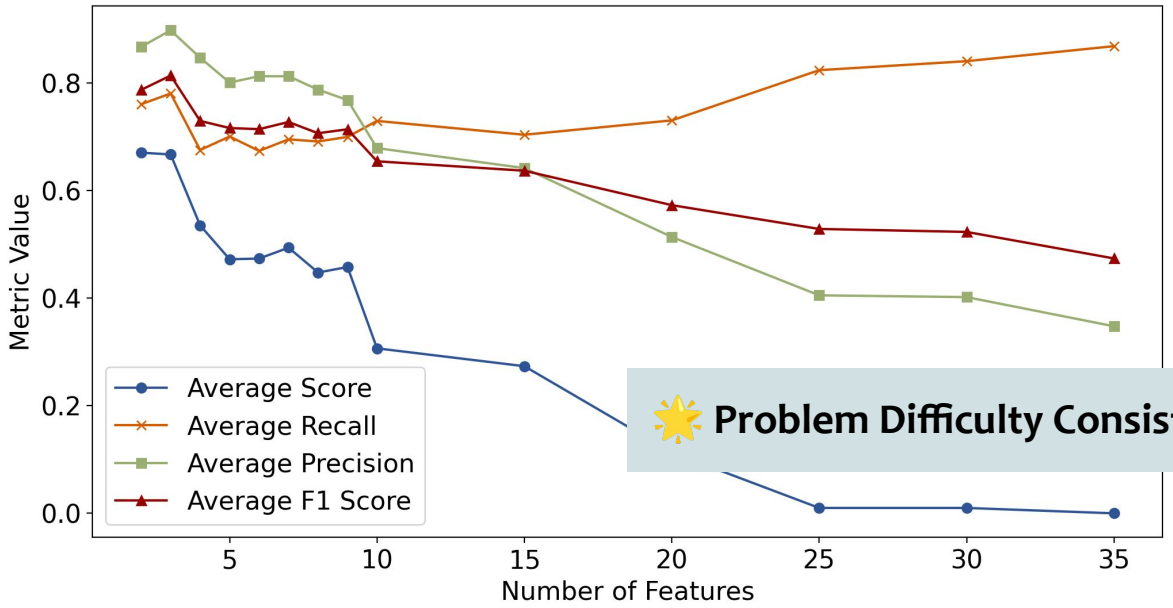

Acc.
1.00
0.75
0.50
0.25
0.00
that_resultative_past_participle
bare_perfect
existential_it
chaining_main_verbs
present_for_neutral_future
superlative_before_matrix_head
definite_abstract
correlative_constructions
corr_conjunction_doubling
negative_inversion
em_obj_pronoun
absolute_reflex
zero_plural_after_quantifier
be_perfect
conditional_were_was
clause_final_really_but
flat_adj_for_adv
inverted_indirect_question
drop_aux_wh
regularized_reflexives
his_he
his_him
their_them
you_ye
those_them
comparative_as_to
comparative_than
do_tense_marker
completive_done
present_perfect_ever
if_would
double_modals
present_modals
past_for_past_participle
participle_past_tense
a_ing
a_participle
negative_concord
aint_have
aint_before_main
never_negator
invariant_tag_can_or_not
invariant_tag_fronted_isnt
generalized_third_person_s
existential_there
drop_aux_be_progressive
drop_aux_have
who_which
who_at
relativizer_where
analytic_whose_relativizer
say_complementizer
existential_you_have
drop_inf_to
to_infinitive
bare_ccomp
clause_final_though_but
subord_conjunction_doubling
null_prepositions
degree_adj_for_adv
too_sub
quotative_like
Feature

1/ Multiple Choice - Multiple Features

Stanford | NLP

Setup2: ask LLM to identify **multiple** linguistic features

- provide LLM a (synthetic) paragraph with **multiple** (2-35) linguistic features



🌟 **Problem Difficulty Consistently Increases!**

2/ Generation From Scratch

Stanford | NLP

Setup3: ask LLM to output linguistic features from scratch

Feature	Example	Reference	Prediction	Specificity	Correctness	BERTScore
Comparative marking only with than	He loves his car than ['more than'] his children.	Using "than" without the preceding comparative word (such as "more" or "less") to indicate a comparison between	the omission of the comparative adjectives 'more' or 'less' before 'than'.	5	5	0.8579
Multiple negation / negative concord	He w				4	0.8886
Ever as marker of experiential perfect	I ever see the movie [I have seen the movie].	Using the word "ever" to indicate an action or experience occurred at some unspecified time in the past, instead of using "have".	Misuse of the adverb 'ever': 'ever' is incorrectly used in affirmative sentences	4	3	0.8582

🌟 **Identifying a linguistic feature from scratch is much more difficult than verifying a given one!**

Linguistic Feature Auto-Discovery

Stanford | NLP

🌟 **More Linguistic Features Possible!**

Conclusion

- **Dynamically aggregation** of non-standard **linguistic features** can better improve LM's multi-dialectal robustness.
- **Linguistic feature discovery** is hard, but **LLMs can help**.

Thank You!