Towards Equitable Natural Language Understanding Systems for Dialectal Cohorts: Debiasing Training Data

Khadige Abboud, Gokmen Oz
Alexa AI, Amazon
abboudk@amazon.com, ogokmen@amazon.de

Abstract

Despite being widely spoken, dialectal variants of languages are frequently considered low in resources due to lack of writing standards and orthographic inconsistencies. As a result, training natural language understanding (NLU) systems relies primarily on standard language resources leading to biased and inequitable NLU technology that underserves dialectal speakers. In this paper, we propose to address this problem through a framework composed of a dialect identification model that is used to obtain targeted training data augmentation for under-represented dialects, in an effort to debias NLU model for dialectal cohorts in NLU systems. We conduct experiments on two dialect rich non-English languages: Arabic and German, using large-scale commercial NLU datasets as well as open-source datasets. Results show that such framework can provide insights on dialect disparity in real-world NLU systems and targeted data argumentation can help narrow the model's performance gap between standard language speakers and dialect speakers.

Keywords: Dialects, Bias, Voice assistants, Low-resource, Data augmentation

1. Introduction

As large language models (LLMs) continue to advance the NLP technology with impressive performance on variety of tasks at the fingertips of millions of people every day, it is important to ensure equity of performance of NLP systems for speakers of different languages and language varieties, i.e., dialects. Prior research has shown that biases exist in these models against certain languages or dialects (Deas et al., 2023; Khondaker et al., 2023) and such bias can start even at the tokenizer level (Petrov et al., 2023) or data quality filters that are applied to data sources prior to the model training (Gururangan et al., 2022) which may put languages of certain demographic groups at an advantage over others. The performance bias is further exacerbated for dialectal varieties.

Dialectal variants pose a unique challenge to language models. Unlike standard languages which have written resources like books, articles, and Wikipedia, the backbone for language models, dialects often lack writing standards and come with considerable differences in phonetics, vocabulary, morphology and syntax compared to their corresponding standard language. These differences are not static and are caused by regional, social, cultural and/or economic factors. Dialectal bias may further marginalize and disenfranchise certain groups by pushing them away from a technology that does not understand their mother tongue dialect. This is also true for voice assistant (VA) systems that are powered by language models to perform natural language understanding (NLU) tasks, such as intent recognition and entity extraction, which are responsible for interpreting users' requests and guiding the VA's response. In households, the common location for VAs, dialectal language is more likely to be used. With the recent advances in conversational AI, dialectal support in VA systems is in fact expected. As these products strive to reach the level of true companionship, they should be able to conduct more natural conversations with users in the language variety of their choice.

In general, biases are inherent in the training data used for pretraining and finetuning (Le et al., 2022; Berthelot et al., 2019; Ng et al., 2020; Gururangan et al., 2022; Le et al., 2023; Garrido Ramas et al., 2022), and dialectal bias is no different. Efforts to address the dialectal disparity have focused on annotated data collection for dialect varieties (Van Der Goot et al., 2021; Plüss et al., 2023; Dogan-Schönberger et al., 2019; Bouamor et al., 2018; El-Haj, 2020; Aepli et al., 2023) or data augmentation through rule-based transformations (Dacon et al., 2022; Ziems et al., 2022, 2023). In a real-world VA system, often samples of user traffic data are annotated and fed back to the model to improve its performance. The dialectal makeup of live traffic data and the information about NLU performance on dialectal cohorts may not be readily available.

In this paper, we build upon the recent works and shed light on dialectal bias from a different angle of a real-world VA system scenario where the dialectal user cohort is unknown. The contributions are as follows:

- We propose a semi-supervised framework for addressing dialectal bias for intent recognition and entity extraction in VA scenarios. The framework consists of a dialect identification model that identifies dialectal cohorts from both training and evaluation data enabling

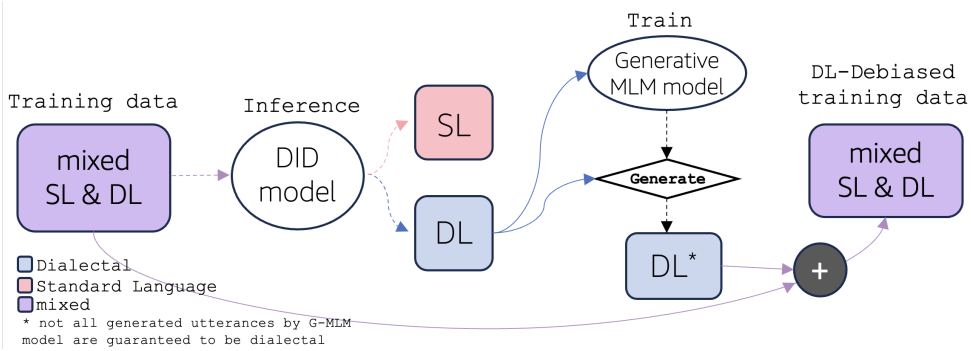

Figure 1: The dialect identification (DID) model is used to segment training data into standard language (SL) and dialectal language (DL). The annotated DL segment of the training data is used to finetune a generative MLM model that is used to augment the training data with more dialectal data to debias the training data of the NLU model.

disparity performance analysis, as well as a generative masked language model (MLM) finetuned on the dialectal subset of the training data for data augmentation and debiasing the training data.

- We conduct experiments on two languages with high dialectal richness, Arabic and German, using the proposed framework. Both languages have large populations of speakers and exhibit significant linguistic diversity, including differences in vocabulary and text.

- We evaluate the performance of dialect-debiased models for intent recognition and entity extraction (I-E) tasks on two large-scale VA dataset as well as two smaller annotated datasets, and show that our proposed dialect-based debiasing framework can help narrow the performance gap for speakers of both German and Arabic dialects.

- We perform further ablation studies to test our design choices and the assumptions made in our experiments. These studies reveal that improvements gained are due to dialect debiasing and are not driven by the volume of training data and these improvements can be seen on independently annotated test sets. Moreover, we conduct additional experiments on LLMs of variable sizes and architectures and demonstrate that our proposed dialect debiasing impact is transferable to larger LLMs.

2. Related Works

Disparity of dialectal cohorts Recent works have highlighted the disparity in dialect performance by state-of-the-art language models and even by LLMs. Models like ChatGPT, GPT-4, GPT-3 underperform on African American English for counterpart language generation task and masked span prediction task, when compared to standard English (Deas et al., 2023). Similar performance gaps have been shown on Arabic dialects (Khondaker et al., 2023). (Petrov et al., 2023) shows that dialectal bias exists even at the tokenizer level, when a monolingual German BERT tokenizer (Scheible et al., 2020) results in better *tokenizer parity*[1] for English than for Swiss German dialects.

Efforts to address the dialectal disparity have mainly taken one of two routes: data augmentation and/or data collection. For the latter, extensive efforts have been put in curating datasets targeted for dialects in several languages, such as Arabic dialects (e.g., Egyptian, Gulf, Levant) (Bouamor et al., 2018; El-Haj, 2020), German (Swiss and South Tyrolean) (Van Der Goot et al., 2021; Plüss et al., 2023; Dogan-Schönberger et al., 2019), and English (African American) (Ziems et al., 2022, 2023). On the data-augmentation front, VALUE (Ziems et al., 2022, 2023) develops a linguistic rule-based framework for augmenting task-specific training data for English dialects and (Srivastava and Chiang, 2023) proposes rule-based transformation and language style disentanglement for African American English dialect.

Debiasing training data Bias is inherent in the training data, mostly due to lack of coverage. Debiasing can be done by unsupervised (Garrido Ramas et al., 2022) or semi-supervised learning (SSL) methods (Le et al., 2022; Berthelot et al., 2019; Ng et al., 2020; Garrido Ramas et al., 2022). SSL methods rely on existing unlabeled data for the target language. When

[1]Tokenizer parity is a metric proposed to assess how fairly a tokenizer treats equivalent sentences in different languages (Petrov et al., 2023).

there is no data, or the information about how much coverage of this data exists is not available, other methods should be sought out. (Dacon et al., 2022; Ziems et al., 2022, 2023) utilize their rule-based transformation to debias the training data by transforming standard seed utterances to different English dialects. (Srivastava and Chiang, 2023) addresses the data scarcity problem of dialects by inducing character-level noise during finetuning BERT to improve cross-lingual transfer learning in zero-shot setting on unseen dialects. (Held et al., 2023) proposes to train adapters with sequence and token level alignment loss between standard English and a target dialect. These task-agnostic dialect adapters can be added before the task-specific standard-English adapter to improve dialect robustness for the target task. These approaches assume prior knowledge of dialectal sub-populations or relies on annotated data which could be absent in many cases. In this paper, we introduce a two-stage dialect-based debiasing framework for real NLU systems. It includes a dialect-identification model to detect dialectal cohorts, and an MLM-based generator trained on identified dialectal data, which is then used to augment and retrain the NLU model.

3. Dialect-based debiasing

Performance bias is inherent in the composition of the pretraining and training data. Even when deliberate effort is made to include and annotate data for the expected target language and dialects, the actual makeup of linguistic and dialectal user cohorts can be different. This is because when an NLU system is up and running online, what and how users decide to speak and ask these systems could be different, particularly when the systems are deployed to a wider regional or even global audience. It is common for NLU systems to improve over time through error resolution or active user feedback, by annotating erroneous utterances and feeding them into the training data to boost model performance. However, it would be preferable to preemptively detect performance bias before erroring out online and causing bad user experiences.

To this end, we propose a dialect-based debiasing approach, illustrated in Figure 1, consisting of two main parts: 1) Dialect cohort extraction through a dialect identification (DID) model trained on a smaller set of dialect tagged dataset, e.g. parallel corpus of transcribed dialectal and standard speech[2], and 2) Synthetic data generation using a masked language model (MLM) finetuned on the identified dialectal training data and trained to generate novel examples by masking part of the annotated training samples (Le et al., 2022). The generative MLM model is then used to augment the training data with dialectal utterances in order to boost the performance of the model for dialectal cohorts.

[2]The dataset used for training the DID model need not be collected from NLU systems (i.e., intent and entity labeling is not needed).

Extracting dialectal sub-population from training data

We train a dialect identification (DID) model as a multi-class classifier that takes utterance text and classifies it as a standard language or one of its dialectal variants. In this work, we focus on dialectal differences detectable from utterance text and not those that can only be detected from phonological differences (speech recognition is out of scope of this paper). For German (**de**), we use SwissDial dataset (Dogan-Schönberger et al., 2019) to finetune a 9-class classifier corresponding to 8 Swiss German dialects and a standard German (**de**). SwissDial dataset contains 26 hours of studio-quality recordings by 8 speakers, each speaking a different German dialect, with both standard German and Swiss German transcripts. For Arabic (**ar**), we build a 3-class classier for modern standard Arabic (MSA) and two popular dialects: Egyptian, and Gulf, using internally collected Amazon Mechanical Turk (mTurk) data. We experimented with different external Arabic song lyrics data labeled with dialect tags (El-Haj, 2020), but we opted for using the simple 3-class model on the mTurk data as it gave a reasonable performance as shown in Table 10 in the Appendix shows the performance using different training data for the **ar**-DID model. We evaluated the DID models on the MADAR benchmark dataset (Bouamor et al., 2018) and xSID dataset (Van Der Goot et al., 2021), for Arabic and German, respectively; as shown in Table 1. Note that for MADAR, we only use the MSA set along with sets from Egyptian speaking regions (Cairo, Alexandria, and Aswan) and Gulf speaking dialects (Jeddah, Riyadh, and Doha).

Table 1: Performance evaluation in terms of accuracy of the dialect identification models for both **de** and **ar** languages, **ar** model is evaluated on the MADAR dataset (Bouamor et al., 2018) and **de** model is evaluated on the xSID dataset (Van Der Goot et al., 2021)

Language	Standard	Dialect
ar	70	77.5
de	87.6	90.4

English (EN):	show|Other all|reference reminders|Other
German (de):	Zeige|Other alle|reference Erinnerungen|Other
South Tyrolean dialect(de-st):	**Zoag|Other olle|reference Erinnerungen|Other**
Swiss German (gsw):	**Du|Other mer|Other au|reference Erinnerige|Other azeige|Other**

Figure 2: Example from xSID data reminder/set_reminder illustrating dialectal differences between South Tyrolean German (de-st) dialect, Swiss German (gsw) dialect, and standard German. The corresponding English translation is included for comparison, dialectal variants are in bold. The slot values for the label *reference* are colored. The intent for this utterance is reminder/show_reminders.

English (EN):	can|Other you turn|Action on|Action the|Other lights|ApplianceCategory
Modern Standard Arabic (ar):	هل|Other يمكنك|Other فتح|Action الإضاءة|ApplianceCategory
Gulf Arabic (ar-GL):	**تقدرين|Other تشتين|Action الليتات|ApplianceCategory**
Egyptian Arabic (ar-EG):	**ممكن |Other تولّعي|Action النور|ApplianceCategory**

Figure 3: Example from **ar** data illustrating dialectal differences between Egyptian Arabic dialect, Gulf Arabic dialect, and Modern Standard Arabic (MSA). The corresponding English translation is included for comparison, dialectal variants are in bold. The slot values for the labels *Action* and *ApplianceCategory* are colored. The intent for this utterance is SmartHome/ApplianceOn. Notice the difference in vocabulary between the dialects and the MSA. The word *Light* is: الإضاءة/alidhAa/ in MSA, الليتات/allaitat/ in Gluf, and النور/alnoor/ in Egyptian.

Dialect-data augmentation

We adopt the generative MLM approach proposed in (Le et al., 2022) for generating novel variants of annotated utterances for intent recognition and entity extraction tasks. The pretrained model is finetuned on MLM task using the NLU training data with the annotation appended to the text to be utilized by the model when filling in masked tokens. During inference, the tokens in the seed utterance are masked with a probability that measures how replaceable the word is, calculated as the number of times pairs of utterances in the seed intent differ only on this word. This probability is calculated from the training data. In this work, the generator model is finetuned using utterances identified as dialectal by our DID model; the same utterance set is then used as the seeds for generation. We set the target to one inference per masked token in the seed utterance. Table 8 in the Appendix lists the total number of generated utterances per dataset and compare it to the original training data. We also look at the uniqueness and novelty of the generated utterances, listed in Table 9 in the Appendix, which shows that at least 72% of the unique utterances generated are not in the baseline training data across the target datasets.

4. Experiments

We evaluate the effectiveness of our proposed debiasing process with three sets of experiments. Two experiments are conducted on real-world VA commercial system, for both German and Arabic; in these systems there is no control over which dialects of the supported language the users choose to speak, i.e., we do not have ground truth labels of the dialect ID for the training or test utterances. The third experiment is conducted on (Van Der Goot et al., 2021) xSID4.0 benchmark; to emulate the target scenario, we shuffle the combined standard and dialectal German training and validation partition and split them 90:10; training:validation partitions and use the DID model to segment them into dialectal and standard cohorts. For xSID experiment, we keep the test sets as is and use their dialect IDs as their ground truth[3].

Setup - For finetuning the DID and NLU models we use DistilmBERT (Sanh et al., 2019) pretrained model which includes Arabic and German. For finetuning the MLM generator we use a pretrained monolingual BERT base models (Antoun et al., 2020) for Arabic and (Staatsbibliothek, 2020) for German, because we want the model to generate utterances in our target language, and since the pretrained models were pretrained for MLM task but on different dataset, the finetuned models would be able to create word substitutions that have not been seen in the rest of the training data and introduced more variability into the augmentation data as pointed out in (Le et al., 2022) Implementation details and hyperparameters are provided in Appendix A.1.

[3]Note that de-dialects do not have train partition in the xSID data. This means that the dialects are low resource in the training data.

Table 2: Absolute SemER scores for baseline and debiased model and relative (% change) SemER between the two models on the German (**de**) xSID (Van Der Goot et al., 2021). Lower values indicate better performance and bold Overall relative % change values indicate improvement in the debiased model with respect to baseline.

Test subset	All			de-standard			de-dialect		
Model / Domain	Baseline	Debiased	%(relative)	Baseline	Debiased	%(relative)	Baseline	Debiased	%(relative)
Alarm	0.28	0.23	-18.26	0.28	0.24	-13.48	0.28	0.22	-20.65
Books	0.18	0.17	-5.72	0.16	0.18	10.13	0.20	0.17	-12.01
CreativeWorks	0.37	0.34	-8.06	0.25	0.26	2.87	0.43	0.38	-11.26
Events	0.53	0.55	3.56	0.38	0.39	3.76	0.61	0.63	3.52
Music	0.46	0.35	-22.40	0.41	0.36	-11.58	0.48	0.35	-26.95
Reminder	0.39	0.36	-8.48	0.40	0.41	2.40	0.39	0.33	-14.10
Reservations	0.38	0.35	-8.75	0.32	0.34	7.37	0.41	0.35	-15.13
Weather	0.31	0.29	-6.52	0.46	0.47	1.03	0.23	0.20	-14.04
Overall	0.36	0.32	**-10.27**	0.36	0.35	**-1.11**	0.36	0.30	**-14.87**

Evaluation datasets - For the main NLU task we use three different datasets for evaluating the intent classification and slot filling task. For German (**de**) language, we evaluate it on (Van Der Goot et al., 2021) xSID4.0 benchmark for intent classification and slot filling which includes standard German and St. Galler-Dütsch dialect (de-gsw) and a very low-resource Austro-Bavarian German dialect, South Tyrolean (de-st): 8 domains, 15 intents, 33 slots and 500 test utterances for each of the standard and dialectal **de** variants. xSID dataset was created as parallel corpus for English utterances extracted at random from benchmarks Snips (Coucke et al., 2018) and the Facebook dataset (Schuster et al., 2019). We also use a large-scale dataset from a real-world VA commercial system of 1.3M utterances spanning 22 domains, 351 intents, and 327 labels.

For Arabic (**ar**) language, we evaluate the NLU tasks on a large-scale dataset consisting of a total 144K **ar** utterances annotated from real-world VA commercial system. This test dataset spans 22 domains, 307 intents, and 246 entity labels. We also evaluate the NLU model on a smaller dataset that spans the same domains which we collect using Mechanical Turk (mTurk) and evaluated by native speakers for the target MSA (24K utterances) and two dialect versions Gulf and Egyptian (24K utterances). Figures 2 and 3 show examples of annotated utterances from German and Arabic datasets that highlight the orthographic differences in dialects when compared to their corresponding standard language.

5. Results and discussion

To evaluate the the performance of the models for I-E tasks we use the semantic error rate (SemER) metric. The semantic error measures how many mistakes are done in entity recognition and slot filling, and is calculated by $SemER = \frac{D+I+S}{C+D+S}$ (Su et al., 2018), where D=deletion, C+D+S I=insertion, S=substitution and C=correct-slots. An intent recognition error is counted as a substitution. In the presented results, *Overall* refers to micro-average, i.e., where all utterances have equal contribution to performance, and *Average* performance is macro-average performance per domain, where each domain has equal weight regardless of its size. The performance is reported as a relative change to the baseline.

Baseline underperforms on dialectal cohorts

We evaluate the effectiveness of our proposed method to reduce the dialectal bias in a real-world VA scenario, in which the makeup of dialectal cohorts is unknown. We use our trained DID model to extract the dialect subgroup of the NLU test set to evaluate the performance of the model (later in the results we also test this assumption with human annotated data in Table 6). To reduce the noise in dialects extraction for test data, we apply the DID model on the carrier phrase of the utterance and not the full utterance text. This is to prevent entities such as song and video names from skewing the inference of the DID model towards one category over the other. We define the carrier phrase as any token labeled as Other or not-a-name related entity (e.g., todo and question entities), for example the utterance "Remind||Other me||Other to||Other drop||todo off||todo rent||todo" will retain all its tokens before DID inference, but an utterance like "Can||Other you||Other play||Other Nickelback||artist" will be stripped of "Nickleback" token before running inference. Note that this carrier phrase extraction can result in no tokens, for example for verbless utterances with only entity names; these utterances are then filtered out from our evaluation datasets. We filter these utterances from the standard and the dialectal evaluation subsets only, but we keep them under the *All* category to track the overall model performance.

Table 3: Relative SemER difference (% change) between baseline and debiased model, on the Arabic (**ar**) and German (**de**) large-scale commercial dataset. Dialect-based debiasing shows improvements on dialectal cohorts extracted by the DID model (de/ar-dialect test subsets). Negative values indicate improvement (Full results in the Appendix).

	Arabic			**German**		
Domain	**All**	**ar-standard**	**ar-dialect**	**All**	**de-standard**	**de-dialect**
Knowledge	-8.5	-12.0	-6.8	-4.33	-2.39	-11.52
Events	-1.67	-7.14	-3.58	3.41	-0.20	9.90
Communication	-0.4	-0.3	-3.3	11.88	5.48	-4.87
SmartHome	-0.7	1.3	-0.3	2.91	1.57	1.50
Music	2.0	2.9	2.0	2.61	-0.65	-0.20
Notifications	-3.1	-7.3	-1.3	-1.84	-4.87	1.20
Weather	10.9	27.7	1.4	-7.30	-7.94	-6.38
Overall	**-0.94**	**-1.04**	**-1.53**	1.56	0.05	**-1.32**

Table 4: Dialectal cohorts disparity with respect to standard language. Relative difference in baseline performance on extracted dialectal cohorts with respect to standard cohorts shows disparity overall for two of the datasets (**ar**) and (**de**) while xSID shows on-par overall performance with average performance disparity. Higher values indicate more performance bias against dialectal cohorts.

	Dataset		
Method	xSID (de)	de	ar
Average	14.26	8.74	-14.53
Overall	0	20.56	14.51

We reserve this cleaning step only for the evaluation dataset and not for the cohort extraction from training. It should be noted that this is only applied to the unlabeled **ar** and **de** evaluation data and not to the xSID data, because it is already annotated with Language/Dialect tags. For each of the experiments, we evaluate the model performance on three test sets: *All*: which contains all the test set utterances unfiltered and uncategorized, *standard*: a subset of *All* set which contains utterances that are labeled or classified as standard language, and *dialect*: a subset of *All* set which contains utterances that are labeled or classified as dialectal variants. Table 4 shows the relative SemER performance between standard language cohorts and dialectal cohorts for the baseline model on the three datasets, with clear bias towards standard language on the three datasets, with the exception of macro-average on **ar**, indicating that dialectal cohorts outperform on some of the smaller domains.

Dialect-based debiasing reduces disparity

Table 2 shows the SemER performance of the debiased models relative to baseline for xSID dataset. Results are averaged on three runs of each baseline and debiased models. Improvements on the xSID dialectal test set (combined St. Galler-Dütsch dialect (de-gsw) and Tyrolean (de-st) dialects) are -14% overall. This improvement is not coming at a cost to the standard German test set, which slightly improve by -1%. We also see improvements on the internal VA datasets, **ar** and **de**, Table 3 shows that debiasing still improves performance of the DID-extracted cohorts but the boost in performance is modest compared to the xSID dataset, this could be because of the feedback loop that is employed in VA systems that learns from samples of traffic. Table 3 lists the overall performance and a few domains for brevity, the full results for all domains are given in the Appendix Tables 11 and 12.

Improvements are not driven by training data volume

To further analyze whether the improvements are due to the augmented data or only because of the added training data volume, we run another set of experiments with simple upsampling to the same size of augmented data of debiased models from previous experiment. Table 5 shows that our method is consistently better than upsampling.

Table 5: Relative SemER difference between baseline and debiased model on the dialectal subset. Dialect-based debiasing consistently outperform random upsampling for dialectal cohorts on the three datasets. Negative values indicate improvement.

	Dataset		
Model	de-dialect (xSID)	de-dialect	ar-dialect
Random upsampling	-6.76	-0.96	16.57
Dialect-based debias	**-14.87**	**-1.32**	**-1.53**

Performance boost still exists on independently annotated test set

Another assumption we questioned is the use of the same DID model for extracting dialectal cohorts from training and evaluation datasets and

whether the same improvements can be seen on an independent set that was not extracted by the same model used to debias the training data. For German language, we evaluated the debiased model on the xSID labeled test sets in the previous section (results in Table 2). For Arabic, we test the same model on internally collected mTurk dataset that is labelled with language/dialect tags[4]. Table 6 shows that similar improvements can be seen on both the DID-model extracted cohorts and the human annotated mTurk set.

Table 6: The dialect-based debiasing for Arabic shows improvements on both dialectal cohorts extracted using the DID model and on mTurk collected data annotated with dialect IDs by native speakers.

	Test subset		
Annotation method	All	ar-standard	ar-dialect
DID model annotated	-0.56	-0.12	-1.02
Human annotated	-0.94	-1.04	-1.53

Dialect-based debiasing improves larger models

To further explore whether our proposed method is effective for LLMs beyond BERT-like models, we conduct ablation studies to evaluate its effectiveness on different model architectures and larger model sizes. We finetune large pretrained language models (LLMs) with decoder-encoder and decoder-only seq2seq architectures of size 5B (Rosenbaum et al., 2022a; FitzGerald et al., 2023), 7B, 20B (Soltan et al., 2022) and 30B[5]. Using these pretrained LLM models, we finetune a baseline model and a debiased model to generate the labeled utterances directly without adding a classification head for the intent recognition and entity extraction tasks. We run these experiments on the xSID dataset (Van Der Goot et al., 2021). Table 7 below shows the semantic error performance of the debiased models relative to their baseline for varying model sizes. Results illustrate that the proposed dialect-based debiasing still provides improvements when coupled with LLMs of different sizes. All four seq2sq LLM models improve on the dialectal test data with larger improvements observed for the 5B and 7B models. Full per-domain results are provided in Table 13 in the Appendix.

[4]This dataset is different than that used to train the DID model and was annotated by language experts.

[5]The 7B and 30B are decoder-only seq2seq models, while the 5B and 20B are encoder-decoder seq2seq models.

Table 7: SemER performance on dialectal test set of xSID dataset (Van Der Goot et al., 2021) for different LLMs. The dialect-based debiasing provides improvements on LLM models of different sizes (size is in terms of number of model parameters in billions (B)).

Model size	baseline	debiased	%(change)
5B	0.41	0.33	**-19.51%**
7B	0.31	0.27	**-12.90%**
20B	0.34	0.32	**-5.88%**
30B	0.24	0.23	**-4.17%**

6. Conclusion

Dialectal varieties of languages pose a significant challenge for language models due to the lack of standardized writing and evolving nature influenced by regional, and cultural contexts. In this paper, we tackle the issue of under-representation of dialects in real-world voice assistant systems. We propose a framework that employs a simple dialect identification (DID) model along with an MLM data generation technique to mitigate biases in the model's training data. With the aid of DID, we extract dialectal cohorts from evaluation data as well, which can shed a light on existing model biases. We conducted experiments on two non-English languages with rich dialectal diversity, Arabic and German. Our results demonstrate that dialect-based debiasing effectively narrow the performance gap for dialectal speakers without adversely impacting standard speaker cohorts. Moreover further experiments and analysis showed that performance gains are consistent across both human-annotated and model-extracted test sets and that these improvements are transferable to varying model sizes and architectures.

Limitations and Ethical Considerations

Limitations - There are a number of limitations for this work. First, our dialect identification is based on utterance text and will not capture phonological differences that do not appear in the text, e.g., the word *schedule* is pronounced as /skedjul/ in Standard American English vs. /shedjul/ in British English, however, it has the same text. Our method can detect dialectal differences from context, vocabulary, spelling ("color" vs. "colour") or syntax.

Secondly, as NLU relies on automatic speech recognition (ASR) output, the accuracy of the dialect id would be limited by ASR errors. Like NLU, ASR also struggles with dialect recognition due to the same challenges discussed in this paper. For example, the utterance "shut up" in Gulf Arabic is تشب/itshub/ and if returned the similar sounding

جب/edjab/, which means "answer" in standard Arabic, DID model will misrecognize the text as standard Arabic and not dialectal. However, in such cases, even intent recognition and slot filling may be different, in our example NLU would return PickUpCall intent and not Silent intent.

Finally, in this work we considered only dialectal vs. standard language. More fine grain studies can be done on each of the dialects supported by these languages, e.g., Swiss dialect can be further categorized based on regions in Switzerland, each with its linguistic differences (Dogan-Schönberger et al., 2019). Similarly, Arabic dialects can be categorized into 25 dialects based on the regions/countries (Bouamor et al., 2018). We leave this for future work. Further future work include exploring other generation approaches that are not limited by masked token generation the one adopted in this paper (Kumar et al., 2022; Rosenbaum et al., 2022b).

Ethical Considerations - The NLU datasets used in this work have been annotated with language experts and mTurk workers, all participants have been compensated and paid for their work. No metadata or personal identifiable information was used in training or evaluating the models.

Acknowledgements

We would like to thank Karolina Owczarzak for her valuable feedback and for her advice and support throughout this research work.

7. Bibliographical References

Khadige Abboud, Olga Golovneva, and Christopher DiPersio. 2022. Cross-lingual transfer for low-resource Arabic language understanding. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 225–237, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Noëmi Aepli, Çağrı Çöltekin, Rob van der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial evaluation campaign 2023. In *Proceedings of the Tenth Workshop on NLP for Similar Languages, Varieties and Dialects*, Dubrovnik, Croatia. ACL.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Jamell Dacon, Haochen Liu, and Jiliang Tang. 2022. Evaluating and mitigating inherent linguistic bias of African American English through inference. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, pages 1442–1454, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nicholas Deas, Jessi Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of african american language bias in natural language generation. *arXiv preprint arXiv:2305.14291*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL HLT*, pages 4171–4186.

Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. 2019. SwissDial: Parallel

multidialectal corpus of spoken swiss german. *ArXiv*, abs/1910.01108.

Mahmoud El-Haj. 2020. Habibi - a multi dialect multi national Arabic song lyrics corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1318–1326, Marseille, France. European Language Resources Association.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Jose Garrido Ramas, Dieu-thu Le, Bei Chen, Manoj Kumar, and Kay Rottmann. 2022. Unsupervised training data re-weighting for natural language understanding with local distribution approximation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 154–160, Abu Dhabi, UAE. ACL.

Suchin Gururangan, Dallas Card, Sarah K Dreier, Emily K Gade, Leroy Z Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A Smith. 2022. Whose language counts as high quality? measuring language ideologies in text data selection. *arXiv preprint arXiv:2201.10474*.

William Held, Caleb Ziems, and Diyi Yang. 2023. TADA : Task agnostic dialect adapters for English. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 813–824, Toronto, Canada. ACL.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Gptaraeval: A comprehensive evaluation of chatgpt on arabic nlp. *arXiv preprint arXiv:2305.14976*.

Manoj Kumar, Yuval Merhav, Haidar Khan, Rahul Gupta, Anna Rumshisky, and Wael Hamza. 2022. Controlled data generation via insertion operations for NLU. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 54–61, Hybrid: Seattle, Washington + Online. ACL.

Dieu-thu Le, Gabriela Hernandez, Bei Chen, and Melanie Bradford. 2023. Reducing cohort bias in natural language understanding systems with targeted self-training scheme. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 552–560, Toronto, Canada. ACL.

Thu Le, Jose Garrido Ramas, Yulia Grishina, and Kay Rottmann. 2022. De-biasing training data distribution using targeted data enrichment techniques. *4th Workshop on Deep Learning Practice and Theory for High-Dimensional Sparse and Imbalanced Data with colocated with KDD 2022*.

Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. 2020. SSMBA: Self-supervised manifold based data augmentation for improving out-of-domain robustness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1268–1283, Online. ACL.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Aleksandar Petrov, Emanuele La Malfa, Philip HS Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. *arXiv preprint arXiv:2305.15425*.

Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, and Mark Cieliebak. 2023. STT4SG-350: A speech corpus for all Swiss German dialect regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772, Toronto, Canada. ACL.

Andy Rosenbaum, Saleh Soltan, Wael Hamza, Yannick Versley, and Markus Boese. 2022a. LINGUIST: Language model instruction tuning to generate annotated utterances for intent classification and slot tagging. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 218–241, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Andy Rosenbaum, Saleh Soltan, Wael Hamza, Yannick Versley, and Markus Boese. 2022b. LINGUIST: Language model instruction tuning

to generate annotated utterances for intent classification and slot tagging. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 218–241, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Raphael Scheible, Fabian Thomczyk, Patric Tippmann, Victor Jaravine, and Martin Boeker. 2020. Gottbert: a pure german language model. *arXiv preprint arXiv:2012.02110*.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. 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)*, pages 3795–3805, Minneapolis, Minnesota. ACL.

Saleh Soltan, Shankar Ananthakrishnan, Jack G. M. FitzGerald, Rahul Gupta, Wael Hamza, Haidar Khan, Charith Peris, Stephen Rawls, Andy Rosenbaum, Anna Rumshisky, Chandana Satya Prakash, Mukund Sridhar, Fabian Triefenbach, Apurv Verma, Gokhan Tur, and Prem Natarajan. 2022. Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model. *arXiv*.

Aarohi Srivastava and David Chiang. 2023. Fine-tuning BERT with character-level noise for zero-shot transfer to dialects and closely-related languages. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 152–162, Dubrovnik, Croatia. ACL.

Bayerische Staatsbibliothek. 2020. German BERT base model. `https://huggingface.co/dbmdz/bert-base-german-uncased"`.

Chengwei Su, Rahul Gupta, Shankar Ananthakrishnan, and Spyros Matsoukas. 2018. A re-ranker scheme for integrating large scale nlu models. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 670–676. IEEE.

Rob Van Der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021. From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2479–2497, Online. ACL.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. VALUE: Understanding dialect disparity in NLU. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3701–3720, Dublin, Ireland. ACL.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. ACL.

8. Language Resource References

Bouamor, Houda and Habash, Nizar and Salameh, Mohammad and Zaghouani, Wajdi and Rambow, Owen and Abdulrahim, Dana and Obeid, Ossama and Khalifa, Salam and Eryani, Fadhl and Erdmann, Alexander and Oflazer, Kemal. 2018. *The MADAR Arabic Dialect Corpus and Lexicon*. Camel labs NYU, Abu Dhabi. PID https://camel.abudhabi.nyu.edu/madar-parallel-corpus/.

Dogan-Schönberger, Pelin and Mäder, Julian and Hofmann, Thomas. 2019. *SwissDial: Parallel Multidialectal Corpus of Spoken Swiss German*. ETH Zurich. PID https://mtc.ethz.ch/publications/open-source/swiss-dial.html.

El-Haj, Mahmoud. 2020. *Habibi - a multi Dialect multi National Arabic Song Lyrics Corpus*. Lancaster University, UK. PID http://ucrel-web.lancaster.ac.uk/habibi/.

Van Der Goot, Rob and Sharaf, Ibrahim and Imankulova, Aizhan and Üstün, Ahmet and Stepanović, Marija and Ramponi, Alan and Khairunnisa, Siti Oryza and Komachi, Mamoru and Plank, Barbara. 2021. *From Masked Language Modeling to Translation: Non-English Auxiliary Tasks Improve Zero-shot Spoken Language Understanding*. ACL. PID https://bitbucket.org/robvanderg/xsid/src/master/.

Table 8: Augmented data sizes using generative MLM model compared to training data volume for each of the experiments on the three datasets

	Train data	Dialectal training data (seed)		Generated data		
Dataset	Size	Size	% of Train data	Size	% of Train data	% of seed dialectal data
xSID	44,507	4,661	10.47	4,541	10.2	97.4
de	16,301,859	7,252,648	44.49	5,156,459	31.6	71.1
ar	1,423,139	847,422	59.55	285,546	20.1	33.7

Table 9: Novelty and uniqueness of the generated data to augment the dialectal training data compared to training data. Comparison is based on utterance text and novel utterances means that they do not appear in the training data. At least 43.2% of total generated utterance are novel and atleast 72.7% of the unique generated utterances are novel across the experiments on the three datasets.

	Generated data		Novel generated data		% Novel data in the generated data	
Dataset	Total size	Unique size	Total size	Unique size	% of Total	% of Unique
xSID	4,541	4,061	3,333	2,952	73.4	72.7
de	5,156,459	2,008,097	2,228,550	1,536,653	43.2	76.5
ar	285,546	98,275	136,833	76,873	47.9	78.2

A. Appendix

A.1. Training details

DID models for both **ar** and **de** are trained by finetuning DistilmBERT (Sanh et al., 2019) with the ADAM optimizer for 40 epochs and batch size of 256 with early stopping based on the f1 scores on the validation set with patience 4 and threshold 0.001.

The I-E models are trained by finetuning DistilmBERT (Sanh et al., 2019) for a joint-task objective with two-layer MLP for the intent recognition task and two-layer MLP plus a CRF layer for the entity extraction (similar to the architecture used in (Abboud et al., 2022)). The models are trained with the ADAM optimizer for 65 epochs (for **de** model we reduce that to 10 epochs due to the large size of the training data) and batch size of 256 with early stopping based on the SemER values on the validation set with patience 4 and threshold 0.001. For further details about the pretrained models, we refer the reader to (Sanh et al., 2019; Antoun et al., 2020; Staatsbibliothek, 2020).

Table 10: Performance evaluation in terms of accuracy of the dialect identification models for **ar** language on the MADAR dataset (Bouamor et al., 2018) for models trained with mTurk collected data, open-source Habibi (El-Haj, 2020) data, or a combination of both.

	Standard	Dialect
Habibi	44.0	32.1
mTurk	**70.0**	**77.5**
mTurk+Habibi	57.0	64.8

A.2. Pretrained models and language resources

DistilmBERT (Sanh et al., 2019) https://huggingface.co/distilbert-base-multilingual-cased is a distilled version of multilingual BERT model is trained on the concatenation of Wikipedia in 104 different languages (including Arabic and German). The model has 6 layers, 768 dimension and 12 head, with a total of 134M parameters.

Both GermanBERT (Staatsbibliothek, 2020) and AraBERT (Antoun et al., 2020) are BERT base models (Devlin et al., 2019) with 12 encoder layers, 768 hidden dimensions, 3072 hidden size, and 12 attention heads, and pretrain for a Masked Language Model (MLM) task. GermanBERT is pretrained on 2.3B tokens from German Wikipedia dump, EU Bookshop corpus, Open Subtitles, CommonCrawl, ParaCrawl and News Crawl and AraBERT is pretrained on 8.6B tokens from Arabic Wikipedia dump, and other Arabic Corpus and news articles.

MADAR dataset (Bouamor et al., 2018) (Bouamor et al., 2018) is a large parallel corpus of 25 Arabic dialects by city in addition to the pre-existing parallel set for English, French and Modern Standard Arabic (MSA), targeting applications of Dialect Identification (DID) and Machine Translation (MT) and not NLU tasks. We use it to evaluate our DID **ar** models. MADAR dataset is not used for training.

xSID dataset (Van Der Goot et al., 2021) (Van Der Goot et al., 2021) was designed for Intent recognition and slot filling in 13 languages from 6 language families, including standard German and Swiss German (St. Galler-Dütsch (**de-gsw**)). The dataset covers 8 domains, 15 intents, 33 slots. The SID4LR shared task, which focuses on Slot and Intent Detection (SID) for digital assistant data (Aepli et al., 2023) expanded the xSID dataset

to cover another low-resource German language variety; the South Tyrolean (**de-st**) dialect. We only use the xSID 4.0 sets from German language: **de**, **de-st** and **de-gsw** for training and testing I-E recognition models.

SwissDial (Dogan-Schönberger et al., 2019) (Dogan-Schönberger et al., 2019) dataset contains 26 hours of studio-quality recordings by 8 speakers, each speaking a different German dialect, with both standard German and Swiss German transcripts originally intended for ASR model training, but we use it for DID training in this paper.

A.3. Augmented data details

We adopt the MLM generation method proposed in (Le et al., 2022). In (Le et al., 2022), the authors aim to debias the training data to match online traffic distribution and use a clustering method to select the target seed utterances. In this work, we set the seed utterances to the training data portion identified as dialectal by the DID model. Table 8 lists the total number of generated utterances per dataset and compare it to the original training data. Note that not every seed utterance would result in a generated utterance as it depends on the carrier phrase (if any) and masked token probability. Further, not all generated utterances are unique or novel. We look at the quality of generated utterances in terms of uniqueness and novelty. Table 9 shows that atleast 72% of the unique utterances generated are novel and are not in the baseline training data. Note that the reason uniqueness of xSID generated data is higher than **ar** and **de** datasets is because the uniqueness of the training data is higher.

A.4. Full per-domain results

Tables 11 and 12 list the performance per domain sorted by size from largest (top) domains to smallest (bottom). Overall, the de-biased model reduces dialectal disparity. Note the varying performance across domains, which could be due to two reasons: 1) the size of the domain; smaller domains are subject to higher fluctuations in performance (e.g., larger relative changes can be seen in small domains like Gallery and News) as opposed to larger domains such as Knowledge and General; and 2) the dialectal variation of utterances within a domain; some domains cover wider variety of utterances (e.g., general Q/A utterances supported under the Knowledge domain) such domains are more likely to benefit from dialectal-debiasing. On the other hand, domains with less utterance diversity such as Music, which supports simple command utterances such as "play a song by Nickelback" are less likely to benefit from dialectal debiasing.

Table 11: Relative SemER difference (% change) between baseline and debiased model, on the Arabic (**ar**) large-scale commercial dataset. Dialect-based debiasing shows improvements on dialectal cohorts extracted by the DID model (ar-dialect test subset). Negative values indicate improvement

	Test subset		
Domain	All	ar-standard	ar-dialect
Music	2.03	2.95	1.95
General	0.57	1.24	-1.99
Knowledge	-8.54	-12.05	-6.79
SmartHome	-0.72	1.30	-0.34
Notifications	-3.12	-7.33	-1.33
Communication	-0.41	-0.28	-3.32
OriginalContent	-3.63	-5.07	-2.86
Events	-1.67	-7.14	-3.58
Weather	10.88	27.67	1.35
Translation	5.48	17.92	-4.81
Movies	2.71	0.77	10.20
Apps	-11.67	-2.70	11.91
Books	-5.18	-4.06	-5.17
Help	4.22	11.33	-0.61
FoodAndHealth	2.77	6.33	2.31
News	34.69	24.40	46.13
Shopping	-0.83	1.57	-1.28
LocalSearch	3.11	0	1.85
Gallery	-9.98	-20.52	-29.99
Car	-15	-26.31	0
Overall	**-0.94**	**-1.04**	**-1.53**

Table 12: Relative SemER difference (% change) between baseline and debiased model, on the German (**de**) large-scale commercial dataset. Dialect-based debiasing shows improvements on dialectal cohorts extracted by the DID model (de-dialect test subset). Negative values indicate improvement

	Test subset		
Domain	All	de-Standard	de-dialect
General	1.58	2.44	2.12
Music	2.61	-0.65	-0.20
SmartHome	2.91	1.55	1.50
Shopping	4.77	3.28	7.66
Knowledge	-4.33	-2.39	-11.52
Notifications	-1.84	-4.87	1.20
Communication	11.88	5.48	-4.87
Weather	-7.30	-7.94	-6.38
LocalSearch	-5.72	-8.54	-9.90
Events	3.41	-0.20	9.90
Books	6.15	1.37	9.54
News	3.11	5.01	0.35
Movies	1.64	3.48	-9.85
OriginalContent	-9.10	-11.04	2.92
FoodAndHealth	-9.53	-14.22	6.36
Apps	-0.37	-3.61	2.27
Help	1.58	3.29	-3
Translation	-10.70	-16.84	-4.91
Car	18.63	27.76	0
Gallery	39.54	33.74	0
Overall	1.56	0.44	**-1.32**

Table 13: Per-domain absolute and relative SemER differences (% change) between baseline and debiased model for different LLM model size, on the dialectal xSID dataset (Van Der Goot et al., 2021). Model sizes are in terms of Billions (B). Negative relative values indicate improvement.

Model size / Domain	5B			7B			20B			30B		
	baseline	debiased	%(change)	baseline	debiased	%(change)	baseline	debiased	%(change)	baseline	debiased	%(change)
Alarm	0.31	0.25	-19.35%	0.23	0.18	-21.74%	0.22	0.23	4.55%	0.14	0.16	14.29%
Books	0.24	0.20	-16.67%	0.17	0.17	0%	0.23	0.20	-13.04%	0.25	0.14	-44.0%
CreativeWorks	0.32	0.23	-28.13%	0.24	0.22	-8.33%	0.42	0.53	26.19%	0.23	0.20	-13.04%
Events	0.66	0.56	-15.15%	0.57	0.50	-12.28%	0.42	0.57	35.71%	0.46	0.44	-4.35%
Music	0.45	0.39	-13.33%	0.41	0.36	-12.20%	0.42	0.35	-16.67%	0.29	0.30	3.45%
Reminder	0.54	0.38	-29.63%	0.35	0.29	-17.14%	0.49	0.32	-34.69%	0.26	0.25	-3.85%
Reservations	0.49	0.46	-6.12%	0.35	0.28	-2 0%	0.43	0.43	0%	0.26	0.25	-3.85%
Weather	0.31	0.24	-22.58%	0.23	0.18	-21.74%	0.18	0.17	-5.56%	0.16	0.17	6.25%
Overall	0.41	0.33	**-19.51%**	0.31	0.27	**-12.90%**	0.34	0.32	**-5.88%**	0.24	0.23	**-4.17%**