Predicting the Target Word of Game-playing Conversations using a Low-Rank Dialect Adapter for Decoder Models

Dipankar Srirag **Aditya Joshi** **Jacob Eisenstein**
University of New South Wales, Sydney Google DeepMind
{d.srirag, aditya.joshi}@unsw.edu.au jeisenstein@google.com

Abstract

Dialect adapters that improve the performance of LLMs for NLU tasks on certain sociolects/dialects/national varieties ('dialects' for the sake of brevity) have been reported for encoder models. In this paper, we extend the idea of dialect adapters to decoder models in our architecture called LoRDD. Using MD-3, a publicly available dataset of word game-playing conversations between dialectal speakers, our task is Target Word Prediction (TWP) from a masked conversation. LoRDD combines task adapters and dialect adapters where the latter employ contrastive learning on pseudo-parallel conversations from MD-3. Our experiments on Indian English and Nigerian English conversations with two models (MISTRAL and GEMMA) demonstrate that LoRDD outperforms four baselines on TWP. Additionally, it significantly reduces the performance gap with American English, narrowing it to 12% and 5.8% for word similarity, and 25% and 4.5% for accuracy, respectively. The focused contribution of LoRDD is in its promise for dialect adaptation of decoder models using TWP, a simplified version of the commonly used next-word prediction task.

1 Introduction

Dialect adaptation of language models refers to approaches that improve their performance for different dialects of a language (Joshi et al., 2025). Past work proposes dialect adaptation for encoder models (Held et al., 2023; Xiao et al., 2023) or encoder-decoder models (Liu et al., 2023). This paper extends it to decoder models, via a novel architecture called **Lo**w-**R**ank **D**ialect robustness for **D**ecoder Models (LoRDD). To demonstrate the effectiveness of LoRDD, we use MD-3 (Eisenstein et al., 2023), a dataset of manually transcribed dialectal dialogues between speakers of either Indian English (en-IN) or Nigerian English (en-NG) or US English (en-US) playing the word-guessing game of taboo[1]. We select MD-3 conversations where the guesser correctly identifies the target word/phrase ('target word' for the sake of brevity) and mask the target word (using [MASK]; as shown in Figure 1). Our task then is to predict the target word in a masked conversation, *i.e.*, target word prediction (TWP). TWP represents a simplified version of next-word generation utilised by decoder models. Since decoder models are adept in tasks involving causal language modeling, TWP is a reasonable task choice. Upon observing that the TWP performances for en-IN and en-NG are lower than those of en-US, the objective of LoRDD is to improve the TWP performances for en-IN and en-NG. LoRDD employs a combination of two LoRA-based (Hu et al., 2022) adapters. The first is a task-specific adapter that uses instruction fine-tuning (Wei et al., 2022) on an augmented set of en-US and en-IN/en-NG conversations. The second is a dialect adapter that uses contrastive learning on a pseudo-parallel corpus between en-US and

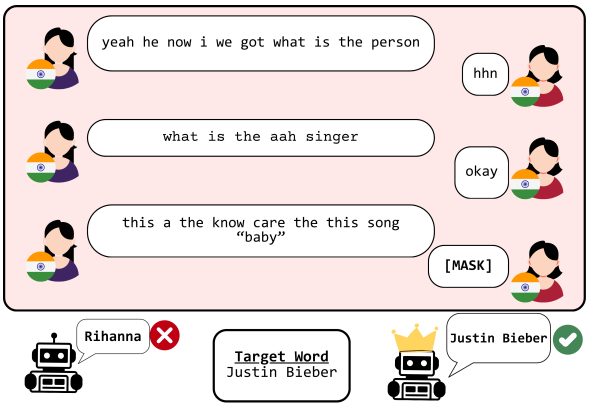

Figure 1: Illustrative example of Target Word Prediction on an en-IN conversation. The inaccurate output from the in-dialect fine-tuned model (left) is corrected by the model trained using LoRDD (right).

[1]In a game of taboo, a describer must get a guesser to guess a target word without using a set of words known as taboo words.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 8–17

en-IN/en-NG conversations about a specific target word. We release the code for training LoRDD adapters on Github.

Our work is novel in two ways: (A) LoRDD is the first methodology for dialect adaptation of decoder models, and outperforms one in-dialect and three cross-dialect baselines, (B) We leverage an existing dataset MD-3 to create a pseudo-parallel corpus of natural dialectal conversations, as opposed to past work that relies on synthetically transformed dialectal corpora.

2 Architecture of LoRDD

The architecture of LoRDD employs two parameter-efficient adapters: task adapter and dialect adapter, as shown in Figure 2.

2.1 Task Adapter

We define $\mathbf{x}$ and $\mathbf{t}$ as lists of tokens in the masked conversation and the target word respectively. For a batched input of N pairs of masked conversations and corresponding target words, we train the task adapters to output the correct target word using maximum likelihood estimation – a standard learning objective for causal language modeling (Jain et al., 2023).

$$\mathcal{L}_{\texttt{Task}} = -\frac{1}{N}\sum_{j=1}^{N}\left\{\sum_{i=|\mathbf{x}^j|+1}^{|\mathbf{x}^j|+|\mathbf{t}^j|}\log p(\mathbf{x}_i^j|\mathbf{x}_{<i}^j)\right\}$$

Here, $\mathbf{x}_{<i}^j = [\mathbf{x}_1^j, \ldots, \mathbf{x}_{i-1}^j]$ denotes the subsequence before $\mathbf{x}_i^j$ and $|\cdot|$ is the number of tokens.

2.2 Dialect Adapter

To train the dialect adapter, we use a pseudo-parallel corpus between en-US and en-IN/en-NG conversations. This corpus consists of both positive and negative pairs of masked conversations. We consider a masked conversation pair as a positive example if both conversations pertain to the same target word, and a negative example if they pertain to a different target word. We then perform contrastive learning between the frozen representation of the masked en-US conversation ($[\text{MASK}]_{\text{US}}$) and the trainable representation of the masked en-IN/en-NG conversation ($[\text{MASK}]_{\text{X}}$), using cosine embedding loss. This allows the adapters to learn from both positive and negative examples present in the pseudo-parallel corpus.

$$\mathcal{L}_{\texttt{Dial}} = \begin{cases} 1 - \text{sim}([\text{MASK}]_{\text{US}}, [\text{MASK}]_{\text{X}}); \text{y} = 1 \\ \max(0, \text{sim}([\text{MASK}]_{\text{US}}, [\text{MASK}]_{\text{X}}) - \text{d}); \text{y} = -1 \end{cases}$$

Here, X represents dialect in focus (either en-IN or en-NG), sim$(\cdot)$ calculates the cosine similarity, 'd' is the margin, and 'y' is the label (1 for a positive example, and -1 otherwise).

In contrast to the task adapter, the dialect adapter is trained to output standard dialect representations for an input text. Hence, LoRDD stacks the task adapter on top of the dialect adapter (as shown in Figure 2), allowing the models to predict the target word as required for TWP.

3 Experiment Setup

We experiment with two open-weight decoder models namely, Mistral-7B-Instruct-v0.2 (MISTRAL; Jiang et al., 2023) and Gemma2-9B-Instruct (GEMMA; Gemma Team, 2024). LoRDD is trained as follows:

- The task adapter is trained by fine-tuning the model for 20 epochs, with a batch size of 32, Paged 8-bit AdamW (Dettmers et al., 2022) as the optimiser and learning rate of 2e-4.
- To train the dialect adapter, we perform contrastive learning for 10 epochs, with a batch size of 8, AdamW as the optimiser, a learning rate of 2e-5, and a margin of 0.25.

We inject adapter matrices at all linear layers, as recommended by Dettmers et al. (2023). Training either adapter for a single experiment takes approx. 25 minutes on an A100 GPU. We compare

Subset	Train	Valid	Test
en-US	62	41	311
en-IN	31	21	160
en-NG	38	25	194
IN-MV	57	39	296
NG-MV	57	39	296
IN-TR	25	17	132

Table 1: Data statistics.

LoRDD with one in-dialect and three cross-dialect baselines. The in-dialect baseline involves fine-tuning a model on the training set of en-IN/en-NG. The cross-dialect baselines are:

en-US Fine-tune the model on train set of en-US.

IN-MV/NG-MV We use Multi-VALUE (Ziems et al., 2023) to transform en-US conversations into en-IN. IN-MV is fine-tuned on these synthetically created conversations.

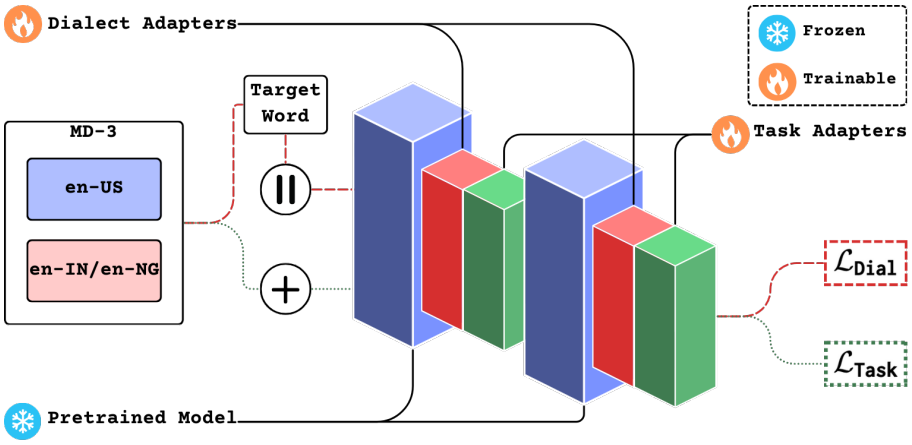

Figure 2: Architecture of LoRDD.

IN-TR We prompt GPT-4 Turbo (OpenAI, 2024) to *transform* en-IN by removing dialectal information, resulting in IN-TR, and use it to fine-tune a model.

Note: We do not perform similar transformations on the en-NG subset due to the high API pricing at the time of writing.

We consider the in-dialect fine-tuned model as a strong baseline, while cross-dialect models are weak baselines. We compare all baselines and LoRDD with in-dialect fine-tuned models on en-US conversations, which serves as our skyline result.

$\|\|_{\text{Corpus}}$	Samples	Positive	Negative
en-US ‖ en-IN	144	11	133
en-US ‖ en-NG	168	13	155
en-US ‖ IN-MV	197	97	100
en-US ‖ NG-MV	197	97	100
en-IN ‖ IN-TR	142	42	100

Table 2: Data statistics of the pseudo-parallel corpus.

Tables 1 and 2 report the statistics of the extended MD-3 dataset and the pseudo-parallel corpus respectively. Additional details including prompt used to create TR-IN and corpus examples are in Appendix A. All evaluations are on the test set of the en-IN or en-NG subsets for the baselines and LoRDD, and on the test set of the en-US dataset for the skyline. We report two metrics: (a) Similarity (average cosine similarity between the Sentence-BERT (Reimers and Gurevych, 2019) embeddings of the reference and generated target word); and (b) Accuracy (the proportion of conversations where the model generates the correct target word).

4 Evaluation

Our results address three questions: (a) What is the current gap in the task performance between en-US and en-IN/en-NG?; (b) How well does LoRDD help bridge the gap?; (c) How essential is each component in LoRDD to bridge the gap?

Table 3 compares the performance of LoRDD with the baselines and the skyline. On the similarity and accuracy, LoRDD achieves average scores of 59.9 and 35.7, respectively, when evaluated on en-IN, and 63.5 and 41.9, respectively, when evaluated on en-NG. On average, LoRDD improves on the performances of the en-IN in-dialect baseline by 13.4% on similarity and 28.1% on accuracy. Similarly, it improves on the en-NG in-dialect baseline by 11.4% on similarity and 33.8% on accuracy.

As expected, the skyline achieves the highest performance for the task. However, LoRDD significantly narrows the initial performance gaps. For en-IN, the gap in similarity is reduced from 27.3% to 12%, and the gap in accuracy is reduced from 64.7% to 25%. For en-NG, the gap in similarity is reduced from 17.9% to 5.8%, and the gap in accuracy is reduced from 43.1% to 4.5%.

Table 4 shows the results from an ablation study that evaluates both adapters in LoRDD. We compare LoRDD with three variants: (a) the dialect adapter trained on other parallel corpora, (b) LoRDD without the dialect adapter, within which we also compare, (c) the task adapters trained on other augmented data. Compared to LoRDD, all other variants report a degradation in their performances. Training the dialect adapter on synthetic parallel corpora (en-US ‖ IN-MV, en-IN ‖ IN-TR and en-US ‖ NG-MV) results in degradation ranging from 1.0 to 2.3 on similarity and 2.5 to 4.8

Method	Training Data	MISTRAL		GEMMA		μ	
		Similarity	Accuracy	Similarity	Accuracy	Similarity	Accuracy
Skyline	en-US	64.7	44.3	69.7	45.3	(0.0) 67.2 (27.3)	(0.0) 44.8 (64.7)
(a) Tested on en-IN							
In-dialect baseline	en-IN	51.0	24.4	54.6	30.0	(27.3) 52.8 (0.0)	(64.7) 27.2 (0.0)
Cross-dialect baseline	en-US	54.6	25.6	61.3	35.0	58.0	30.3
	IN-MV	52.4	24.4	58.2	30.0	55.3	27.2
	IN-TR	50.4	24.3	53.0	26.9	52.7	25.6
LoRDD	en-US + en-IN	**55.9**	**30.0**	**63.9**	**41.3**	(12.0) **59.9** (13.4)	(25.0) **35.7** (28.1)
(b) Tested on en-NG							
In-dialect baseline	en-NG	53.0	27.2	60.9	35.3	(17.9) 57.0 (0.0)	(43.1) 31.3 (0.0)
Cross-dialect baseline	en-US	58.9	31.4	62.8	40.7	60.9	36.1
	NG-MV	55.7	28.4	61.4	38.6	58.9	33.5
LoRDD	en-US + en-NG	**62.4**	**40.5**	**64.5**	**43.2**	(5.8) **63.5** (11.4)	(4.5) **41.9** (33.8)

Table 3: Performance comparison between the skyline, baselines and LoRDD on TWP. For each model, we report Similarity and Accuracy when tested on (a) en-IN and (b) en-NG. μ is the average of the metrics across both evaluation models. LoRDD (represented in **bold**) improves the performance on all baselines. The percentage improvement over the in-dialect baseline and the percentage degradation compared to the skyline are shown in (number) and (number) respectively.

Method	Training Data	$\Vert_{\text{Corpus}}$	MISTRAL		GEMMA		μ	
			Similarity	Accuracy	Similarity	Accuracy	Similarity	Accuracy
(a) Tested on en-IN								
LoRDD	en-US + en-IN	en-US ‖ en-IN	**55.9**	**30.0**	**63.9**	**41.3**	**59.9**	**35.7**
$\leftrightarrow \Vert_{\text{Corpus}}$	en-US + en-IN	en-US ‖ IN-MV	55.6	28.1	62.0	37.5	58.8 (1.1)	32.8 (2.9)
	en-US + en-IN	en-IN ‖ IN-TR	54.9	27.5	62.8	38.8	58.9 (1.0)	33.2 (2.5)
$-\mathcal{L}_{\texttt{Dial}}$	en-US + en-IN	Not Used	54.4	26.9	62.3	37.5	58.4 (1.5)	32.2 (3.5)
	en-IN + IN-MV		51.6	23.1	57.1	31.9	54.4 (5.5)	27.5 (8.2)
	en-IN + IN-TR		44.8	18.1	57.5	28.8	51.2 (8.7)	23.5 (12.2)
(b) Tested on en-NG								
LoRDD	en-US + en-NG	en-US ‖ en-NG	**62.4**	**40.5**	**64.5**	**43.2**	**63.5**	**41.9**
$\leftrightarrow \Vert_{\text{Corpus}}$	en-US + en-NG	en-US ‖ NG-MV	60.4	35.6	61.9	38.5	61.2 (2.3)	37.1 (4.8)
$-\mathcal{L}_{\texttt{Dial}}$	en-US + en-NG	Not Used	61.3	39.7	62.4	38.1	61.9 (1.6)	38.9 (3.0)
	en-IN + NG-MV		58.6	33.6	60.7	33.1	59.7 (3.8)	33.4 (8.5)

Table 4: Ablation on LoRDD based on parallel corpus ($\leftrightarrow \Vert_{\text{Corpus}}$), dialect adapter ($\mathcal{L}_{\texttt{Dial}}$) and data augmentation. For each model, we report Similarity and Accuracy when tested on (a) en-IN and (b) en-NG. The best performance is shown in **bold**. μ is the average of the metrics across both models. The degradation on the ablations compared to LoRDD is shown in (number).

on accuracy. Removing the dialect adapter results in a further degradation ranging from 1.5 to 8.7 on similarity and 3.0 to 12.2 on accuracy. The worst-performing variants are the models that only train the task adapter on synthetically augmented data (en-US + IN-MV, en-IN + IN-TR and en-IN + NG-MV). While the degraded performances of these models show the importance of the dialect adapter, the lower performances on variants involving synthetic conversations further solidify the use of natural conversations in LoRDD. We provide additional results, such as ablations on proportion of conversations in augmented data, in Appendix B.

Finally, we manually analyse erroneous en-IN instances from LoRDD, and categorise them into types of en-IN dialect features given by Lange (2012) and Demszky et al. (2021). Figure 3 shows that EXTRANEOUS ARTICLE (*"It's a one word"*) is the most common feature associated with these conversations. The definitions of all identified dialect features with examples are in Table 5.

Note: We do not perform error analysis for en-NG instances due to lack of similar labelled features for the dialect.

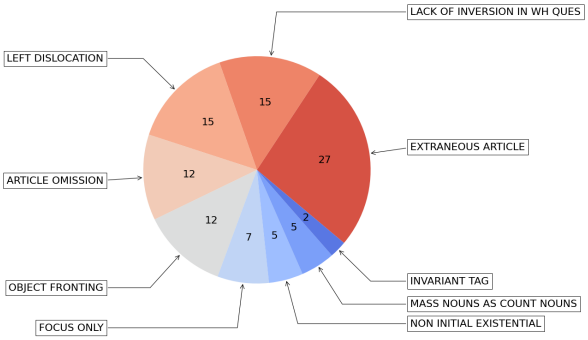

Figure 3: Percentage count of dialect features in erroneous instances from LoRDD.

Feature	Example
EXTRANEOUS ARTICLE	*you can combine the both the words*
LACK OF INVERSION IN WH-QUESTIONS	*what we can see in the rivers?*
LEFT DISLOCATION	*If we have a five sides, what do we call that?*
ARTICLE OMISSION	*I'll explain you (the) second word*
OBJECT FRONTING	*some towers type it will be*
FOCUS *only*	*I'm trying to explain that only*
NON-INITIAL EXISTENTIAL	*brand names also there*
MASS NOUNS AS COUNT NOUNS	*How the womens will be?*
INVARIANT TAG	*put them on some type of wire no?*

Table 5: Dialect features identified in erroneously labelled en-IN conversations with the corresponding examples.

5 Related Work

Language technologies need to be equitable to dialects/sociolects/national varieties (Joshi et al., 2025; Blodgett et al., 2020). Dialect adaptation involves strategies to improve the performance of non-mainstream dialects. These strategies range from introducing dialectal information at the pretraining phase (Sun et al., 2023) to adapter-based approaches. Adapters are explored to be viable and efficient in improving dialect robustness (Liu et al., 2023) or cross-lingual transfer (Pfeiffer et al., 2020). In particular, we derive from this line of work by training a low-rank dialect adapter like Xiao et al. (2023) using a contrastive learning objective like Held et al. (2023). While past approaches adapt encoder models, we distinguish ourselves by proposing LoRDD as an architecture to adapt decoder models. Similarly, past work uses frameworks like VALUE (Ziems et al., 2022) and Multi-VALUE (Ziems et al., 2023) to create synthetic dialectal variants of standard US English benchmarks. In contrast, we use a pseudo-parallel corpus of naturally occurring dialectal conversations from MD-3 (Eisenstein et al., 2023). Our task of target word prediction is closely similar to Chalamalasetti et al. (2023), who generate word game conversations using LLMs and evaluate their ability to predict the target word. Target word prediction is also utilised by Srirag et al. (2025), who evaluate dialect-robustness of language models using masked MD-3 conversations. Finally, our cross-dialect baselines on corpora created using Multi-VALUE and GPT-4 discuss the shortcomings of synthetic datasets for dialect adaptation for dialogues, as also noted in Faisal et al. (2024).

6 Conclusion

This paper focused on a simplistic causal language modeling task, called target word prediction, using masked game-playing conversations between two dialectal speakers of English (en-US, en-IN and en-NG). The task was to predict the target word from a masked conversation. From our initial experiments with fine-tuned decoder models, the in-dialect baseline (en-IN and en-NG) reported a performance degradation on TWP, when compared with the skyline (en-US). To address the gap in the case of en-IN and en-NG, we proposed LoRDD as a novel architecture using low-rank adapters. LoRDD extends past work in dialect adaptation for encoder models to decoder models by employing contrastive learning via a pseudo-parallel corpus of real conversations. LoRDD outperformed one in-dialect baseline and three cross-dialect baselines, while also bridging the gap with the skyline to 12% (down from 27.3%) and 25% (down from 64.7%) on similarity and accuracy respectively for en-IN. For en-NG, the gap is reduced to 5.8% (down from 17.9%) on similarity and 4.5% (down from 43.1%) on accuracy. Through ablation tests on LoRDD, we validated the effectiveness of its components.

Although TWP works with a restricted dataset and utilises turn-based dialogue, LoRDD sets up the promise for dialect adaptation of decoder models. Our error analysis also highlights the scope for future improvement. A potential future work is to evaluate LoRDD on other causal language modeling tasks, including seq2seq tasks, and other dialects. Similarly, an extension to LoRDD would eliminate the requirement of naturally occurring conversations in multiple dialects.

Limitations

While previous approaches have proposed dialect adapters as task-agnostic, our study does not make the same claim. We use target word prediction as

the task of predicting the last word of a conversation which was the word that the described was attempting to convey to the guesser. This task is a simplistic version of causal language modeling. However, we do not verify that LoRDD works for causal language modeling because there is no suitable parallel dataset of turn-aligned conversations, to the best of our knowledge. Held et al. (2023) use bottleneck adapters based on their ability for cross-lingual transfer, but we do not explore these types of adapters due to the lack of support for our choice of models at the time of writing the paper. The choice of en-IN and en-NG as the dialects of interest is solely based on the availability of the dataset.

Ethics Statement

We use a publicly available dataset of conversations consisting of human players engaged in a game of taboo. The topics discussed in the dataset are fairly general and are unlikely to cause distress. One of the authors of the paper performed the error analysis. The synthetic conversation created using GPT-4 may contain biased output, arising due to the properties of the model. We do not expect any reasonably significant risks arising as a result of the project.

References

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. 2023. clembench: Using game play to evaluate chat-optimized language models as conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11174–11219, Singapore. Association for Computational Linguistics.

Dorottya Demszky, Devyani Sharma, Jonathan Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2021. Learning to recognize dialect features. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2315–2338, Online. Association for Computational Linguistics.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit optimizers via block-wise quantization. In *International Conference on Learning Representations*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jacob Eisenstein, Vinodkumar Prabhakaran, Clara Rivera, Dorottya Demszky, and Devyani Sharma. 2023. Md3: The multi-dialect dataset of dialogues. In *INTERSPEECH*.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: An NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

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. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Nihal Jain, Dejiao Zhang, Wasi Uddin Ahmad, Zijian Wang, Feng Nan, Xiaopeng Li, Ming Tan, Ramesh Nallapati, Baishakhi Ray, Parminder Bhatia, Xiaofei Ma, and Bing Xiang. 2023. ContraCLM: Contrastive learning for causal language model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6436–6459, Toronto, Canada. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. Natural language processing for dialects of a language: A survey. *ACM Comput. Surv.* Just Accepted.

C. Lange. 2012. *The Syntax of Spoken Indian English*. Varieties of English around the world. John Benjamins Publishing Company.

Yanchen Liu, William Held, and Diyi Yang. 2023. DADA: Dialect adaptation via dynamic aggregation of linguistic rules. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13776–13793, Singapore. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. 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)*, pages 7654–7673.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Dipankar Srirag, Nihar Ranjan Sahoo, and Aditya Joshi. 2025. Evaluating dialect robustness of language models via conversation understanding. In *Proceedings of the Second Workshop on Scaling Up Multilingual & Multi-Cultural Evaluation*, pages 24–38, Abu Dhabi. Association for Computational Linguistics.

Jiao Sun, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein, and Sebastian Gehrmann. 2023. Dialect-robust evaluation of generated text. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6010–6028, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Zedian Xiao, William Held, Yanchen Liu, and Diyi Yang. 2023. Task-agnostic low-rank adapters for unseen English dialects. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7857–7870, Singapore. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-value: A framework for cross-dialectal english nlp. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

A Dataset Construction

Table 6 describes the example conversations from en-IN and en-US subsets along with their respective transformed IN-TR and IN-MV conversations. We utilise the following prompt used in the evaluation study by Srirag et al. (2025) to create IN-TR.

> *'Normalise the conversation. Remove all exaggerations and dialectal information. Return a neutral response.'*

The conversations are then masked by replacing the target word with the [MASK] token and pruning the rest of the conversation, as described in the Table 7.

en-IN	IN-TR
Describer: (Uh). What do you call **if we, what will be there** in the water? **Guesser:** Fish(es) **Describer:** Who **will catch that**? **Guesser:** Fisherman.	**Describer:** (∅) What do you call **the creatures** in the water? **Guesser:** Fish(∅). **Describer:** Who **catches them**? **Guesser:** Fishermen.
en-US	**IN-MV**
Describer: Perfect. Oh! (We) earn this. We go to our jobs. **Guesser:** Money	**Describer:** Perfect. Oh! (∅) **[are]** earn**[ing]** this. We **[are]** go**[ing]** to our jobs. **Guesser:** Money

Table 6: Example *transformations* of en-IN to IN-TR, and en-US to IN-MV. We utilise GPT-4 Turbo to generate IN-TR, and Multi-VALUE to create IN-MV. The text in parentheses refers to the omission/removal of certain filler and exaggerated words, and the text such as **this**, refers to the words or sentences that were rephrased to convey the original meaning, and the text such as **[this]**, refers to the dialectal features added using Multi-VALUE.

Table 8 describes examples from the pseudo-parallel corpus: en-US || en-IN. The conversations in a positive pair, while dissimilar in the syntax of the conversation, pertain to the same target word. For example, the conversation pair labelled as 'positive' in the Table 8 describe the same target word– *Washing Machine*. The conversation pair labelled as 'negative' describe different target words; the en-US conversation describes *Justin Bieber*, while en-IN conversation describes *Washing Machine*.

B Additional Ablations

We conducted additional ablation studies on LoRDD to address the following question: Can the performance improvement of LoRDD be attributed to the increased training data from data augmentation?

Table 9 compares the performance of the proposed combination of LoRDD with variations that exclude data augmentation. Training the task adapter solely on en-IN results in significantly lower performance, with similarity scores dropping by 5.9 to 7.0 and accuracy scores decreasing by 8.2 to 9.7.

Table 10 examines the effect of varying the proportion of en-US conversations in the augmented training data (en-US + en-IN). The best performance is observed when LoRDD is trained with augmented data containing only 50% en-US conversations. While this configuration outperforms the proposed full-proportion combination, determining the optimal proportions is challenging and limits generalisability across models. More particularly, Table 10 also reveals that MISTRAL is highly sensitive to such changes in the training data composition, whereas GEMMA is more robust.

These ablation results, combined with the findings in Table 4, further reinforce our proposed methodology. Specifically, training the task adapter on fully proportioned augmented data (en-IN + en-US) and the dialect adapter on a parallel corpus constructed from natural conversations (en-US || en-IN) proves to be a more effective and generalisable approach.

Target Word	en-IN	Masked en-IN
Fisherman	Describer: Uh. What do you call if we, what will be there in the water? Guesser: Fishes Describer: Who will catch that? Guesser: **Fisherman**.	Describer: Uh. What do you call if we, what will be there in the water? Guesser: Fishes Describer: Who will catch that? Guesser: **[MASK]**
Target Word	**en-US**	**Masked en-US**
Planet	Describer: These are hard words. um Okay. So there's. the Sun and the Moon and all the rest of them. Guesser: And all the **planet**s? **(**Describer: Yes.**)**	Describer: These are hard words. um Okay. So there's. the Sun and the Moon and all the rest of them. Guesser: **[MASK]**

Table 7: Masking conversations from the extended MD-3. The text such as **this** represents the target word utterance by the guesser which is masked (represented by, **[MASK]** in the final version of the conversation. The rest of the original conversation is pruned as represented text in parentheses.

Label	en-US	en-IN
Positive	**Describer:** Good job. Okay. Um. How we. How we clean our clothes. **Guesser:** [MASK]	**Describer:** Yeah here I got a thing uh which most of us daily use that to wash our clothes. **Guesser:** [MASK]
Negative	**Describer:** this. What? All right all right so. **Guesser:** What? **Describer:** Uh this uh this young man. um is a very well-known singer. who was kind of a heart-throb. Hm he I mean he's still active but like 10 years ago like all of the girls were crazy about this guy. **Guesser:** [MASK]	**Describer:** Yeah here I got a thing uh which most of us daily use that to wash our clothes. **Guesser:** [MASK]

Table 8: Example conversation pairs from the pseudo-parallel corpus: en-US || en-IN. A positive example contains conversations describing the same target word, while the negative example contains conversations pertaining to two different target words.

Method	Training Data	$\|\|_{\text{Corpus}}$	MISTRAL		GEMMA		μ	
			Similarity	Accuracy	Similarity	Accuracy	Similarity	Accuracy
LoRDD	en-US + en-IN	en-US || en-IN	**55.9**	**30.0**	**63.9**	**41.3**	**59.9**	**35.7**
$\leftrightarrow \|\|_{\text{Corpus}}$	en-IN (No Augmentation)	en-US || en-IN	52.0	23.1	53.7	28.8	52.9 (7.0)	26.0 (9.7)
		en-IN || IN-TR	52.0	23.8	54.1	28.8	53.0 (6.9)	26.3 (9.4)
		en-US || IN-MV	53.3	25.0	54.6	30.0	54.0 (5.9)	27.5 (8.2)

Table 9: Ablation on LoRDD based on parallel corpus ($\leftrightarrow \|\|_{\text{Corpus}}$) and data augmentation. For each model, we report Similarity and Accuracy when tested on en-IN. The best performance is shown in **bold**. μ is the average of the metrics across both models. The degradation on the ablations compared to LoRDD is shown in (number).

Method	$\mathbb{I}_{\text{Corpus}}$	% of en-US	MISTRAL		GEMMA		μ	
			Similarity	Accuracy	Similarity	Accuracy	Similarity	Accuracy
LoRDD	en-US ‖ en-IN	0%	52.0	23.1	53.7	28.8	52.9	26.0
		25%	53.8	31.9	61.2	35.4	57.5	33.7
		50%	**58.8**	**33.8**	**64.1**	**41.8**	**61.5**	**37.8**
		75%	54.6	30.6	63.4	40.8	59.0	35.7
		100%*	55.9*	30.0*	63.9*	41.3*	59.9*	35.7*
$-\mathcal{L}_{\texttt{Dial}}$	Not Used	0%	51.0	24.4	54.6	30.0	52.8	27.2
		25%	52.0	29.4	60.5	34.4	56.3	31.9
		50%	55.3	29.4	61.4	35.6	58.4	32.2
		75%	52.5	27.5	61.6	35.6	57.1	31.6
		100%	54.4	26.9	62.3	37.5	58.4	32.2

Table 10: Ablation on LoRDD based on dialect adapter ($\mathcal{L}_{\texttt{Dial}}$) and proportion of en-US conversations in augmented data (en-US + en-IN). For each model, we report Similarity and Accuracy when tested on en-IN. The best performance is shown in **bold**, and the proposed combination is represented by number*. μ is the average of the metrics across both models.