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

Dipankar Srirag

d.srirag@unsw.edu.au

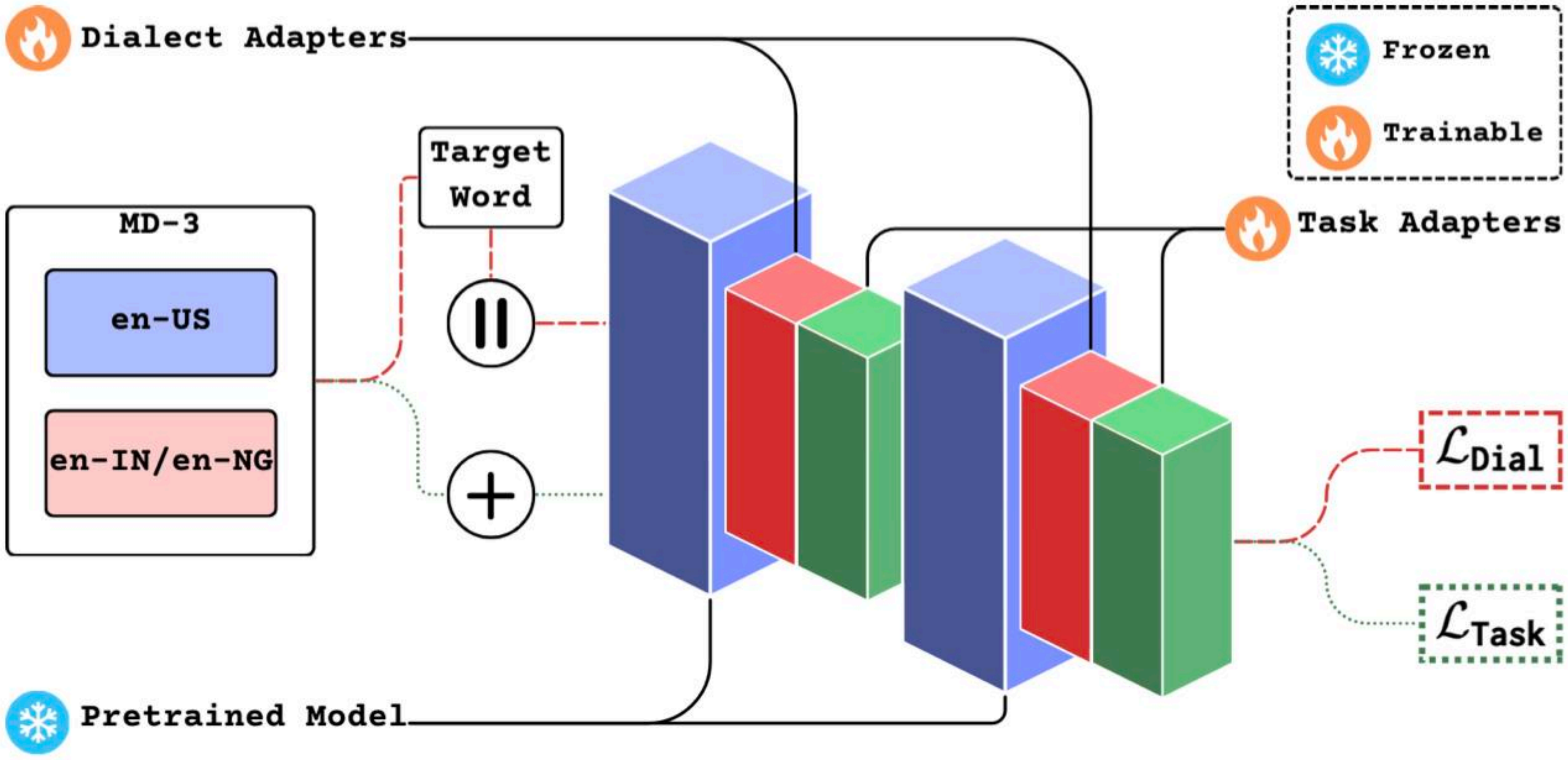
Aditya Joshi

Jacob Eisenstein

2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)

1 May 2025

TL;DR



We train dialect adapters using contrastive learning objective.

Our Architecture is called LoRDD
(Low-Rank Dialect robustness for Decoder Models)

Task Definition

Topic Classification as Target Word Prediction

- Given a masked dialogue between dialectal speakers playing a game of taboo, predict the target word.

Dataset

- MD-3 (**Eisenstein et al., 2023**)

Dialects

- American English
- Indian English
- Nigerian English

en-US
en-IN
en-NG

Eisenstein, J., Prabhakaran, V., Rivera, C., Demszky, D., Sharma, D. (2023) MD3: The Multi-Dialect Dataset of Dialogues. Proc. Interspeech 2023, 4059-4063, doi: 10.21437/Interspeech.2023-2150

Motivation

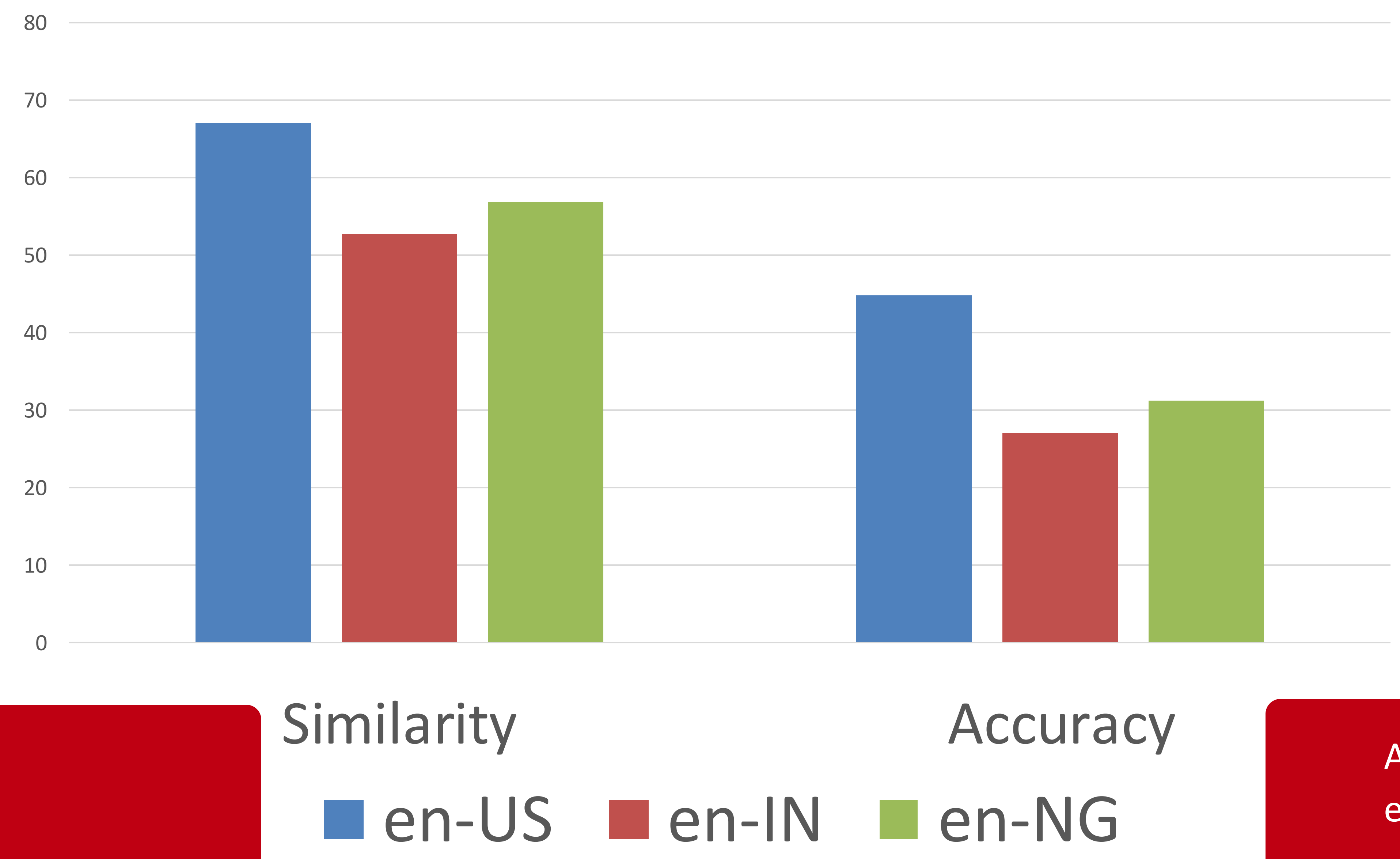
There exists a performance gap between American English and other dialects of English for several NLP tasks
(**Joshi et al., 2025**).

The failure of language technology to cope with dialect differences may create allocational harms that reinforce social hierarchies
(**Blodgett et al., 2020**).

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. 57, 6, Article 149 (June 2025), 37 pages. https://doi.org/10.1145/3712060

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.

Why dialect adapters?



Average gap between en-IN and en-US:

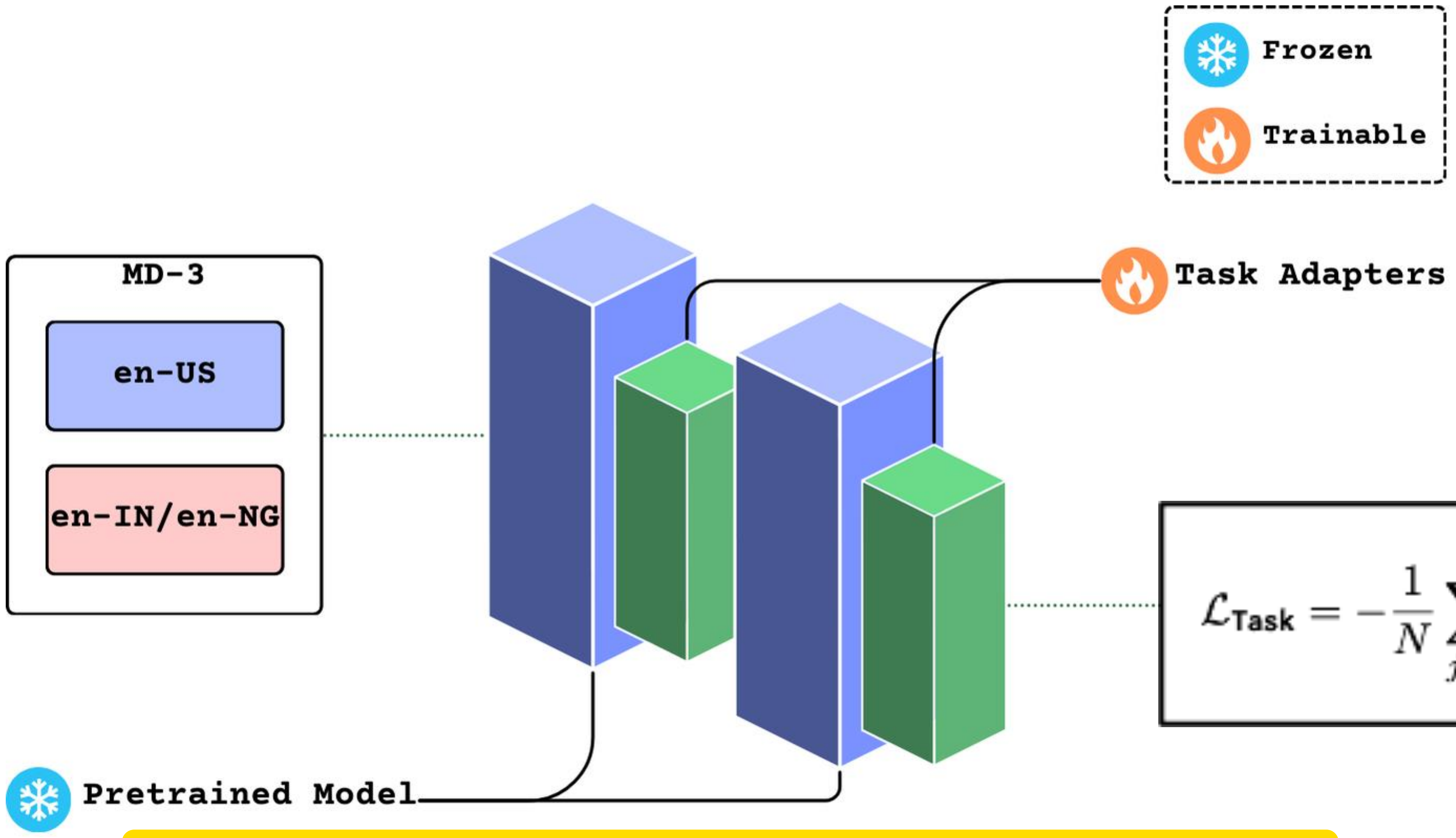
- **27.3%** for Similarity
- **64.7%** for Accuracy

Trained and tested on same dialect.
Averaged across mistral and gemma.

Average gap between en-NG and en-US:

- **17.9%** for Similarity
- **43.1%** for Accuracy

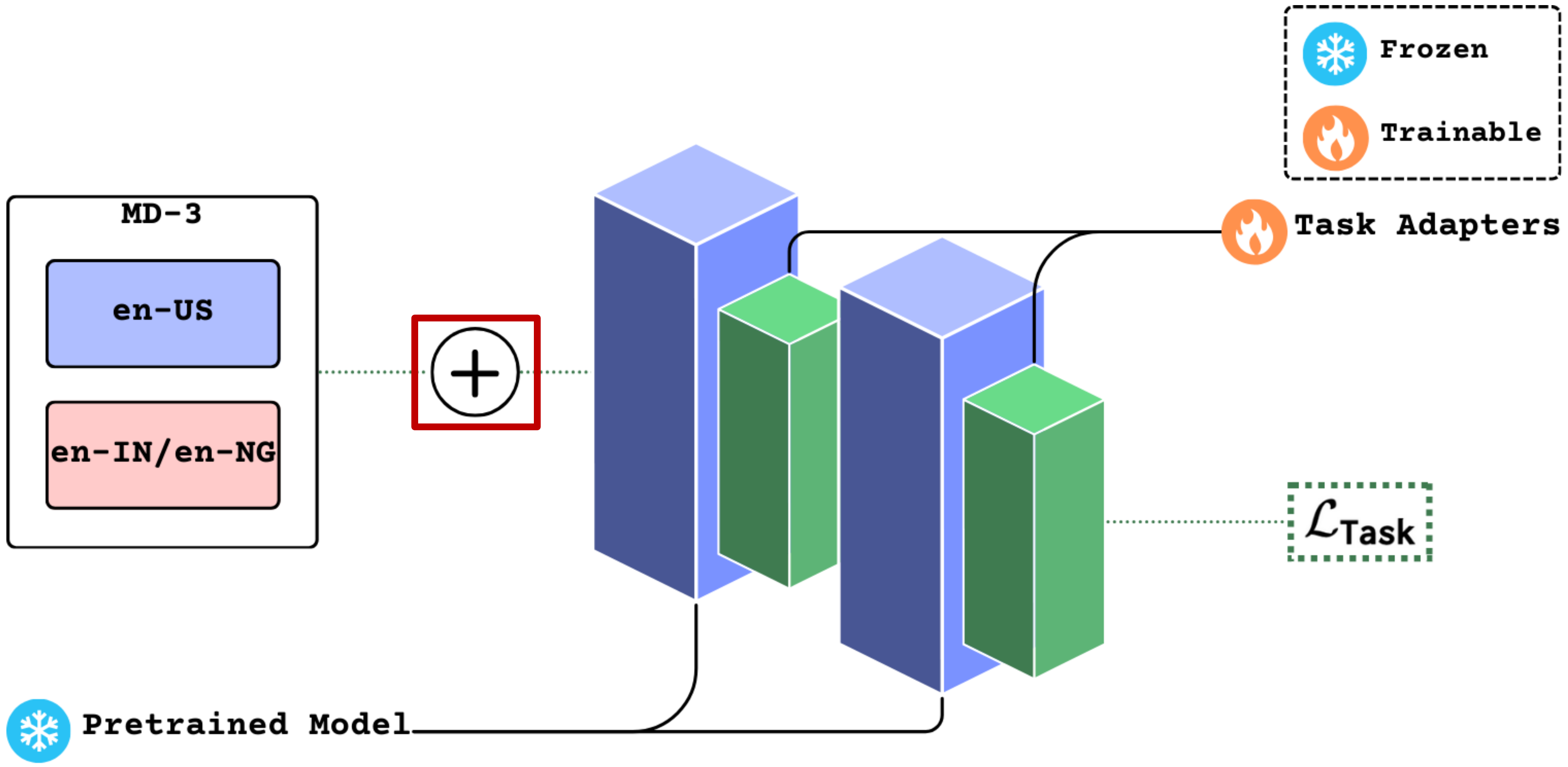
Task Adapters



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

LoRA as the task adapter for **Target Word Prediction**

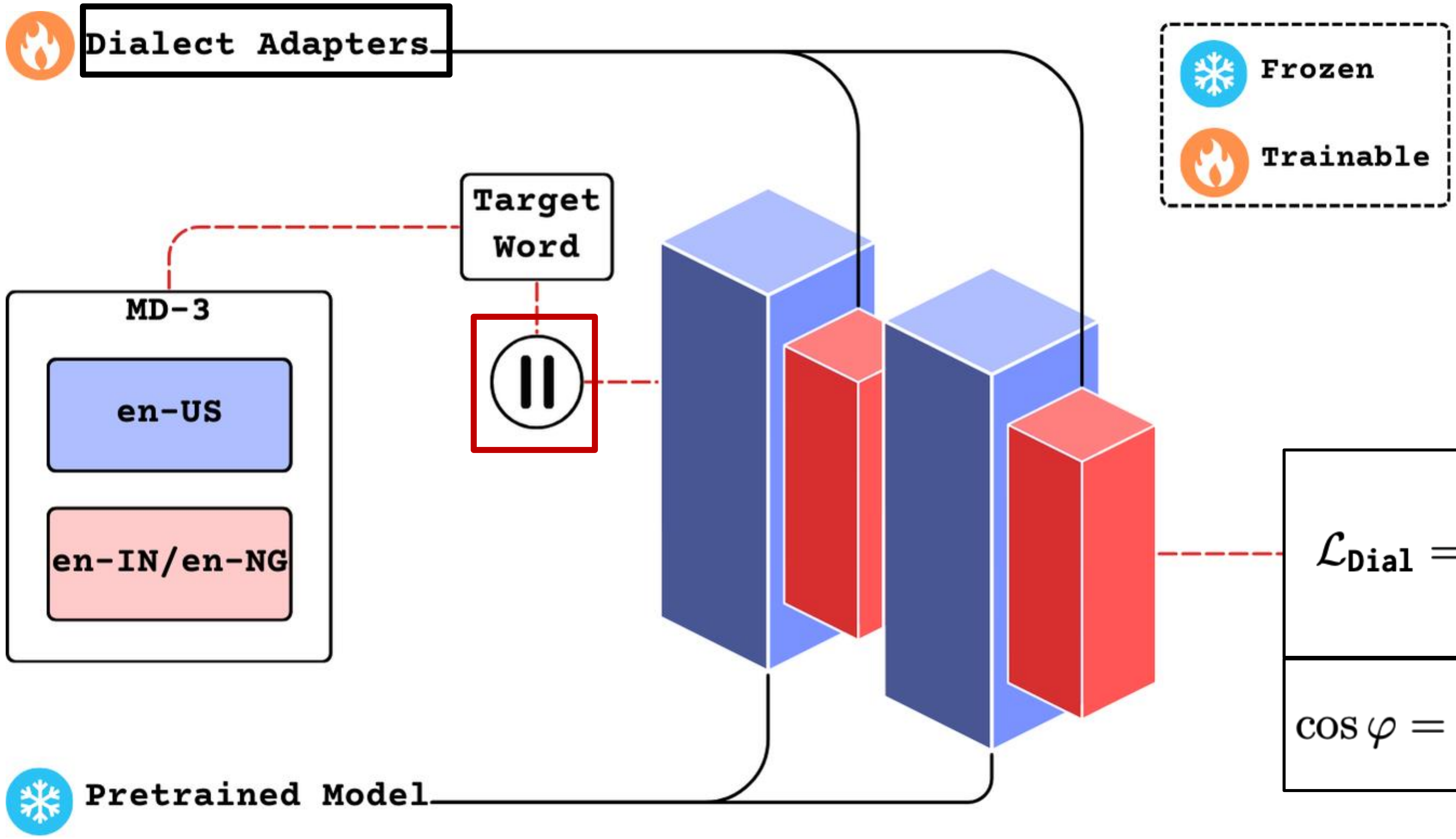
+Data Augmentation



Augment training data:
- en-US
- en-IN/en-NG

+Contrastive Loss

LoRA as **dialect adapter**



$$\mathcal{L}_{\texttt{Dial}} = \begin{cases} 1 - \cos\varphi & ; y = 1 \\ \max(0, \cos\varphi - d) & ; y = \text{-}1 \end{cases}$$

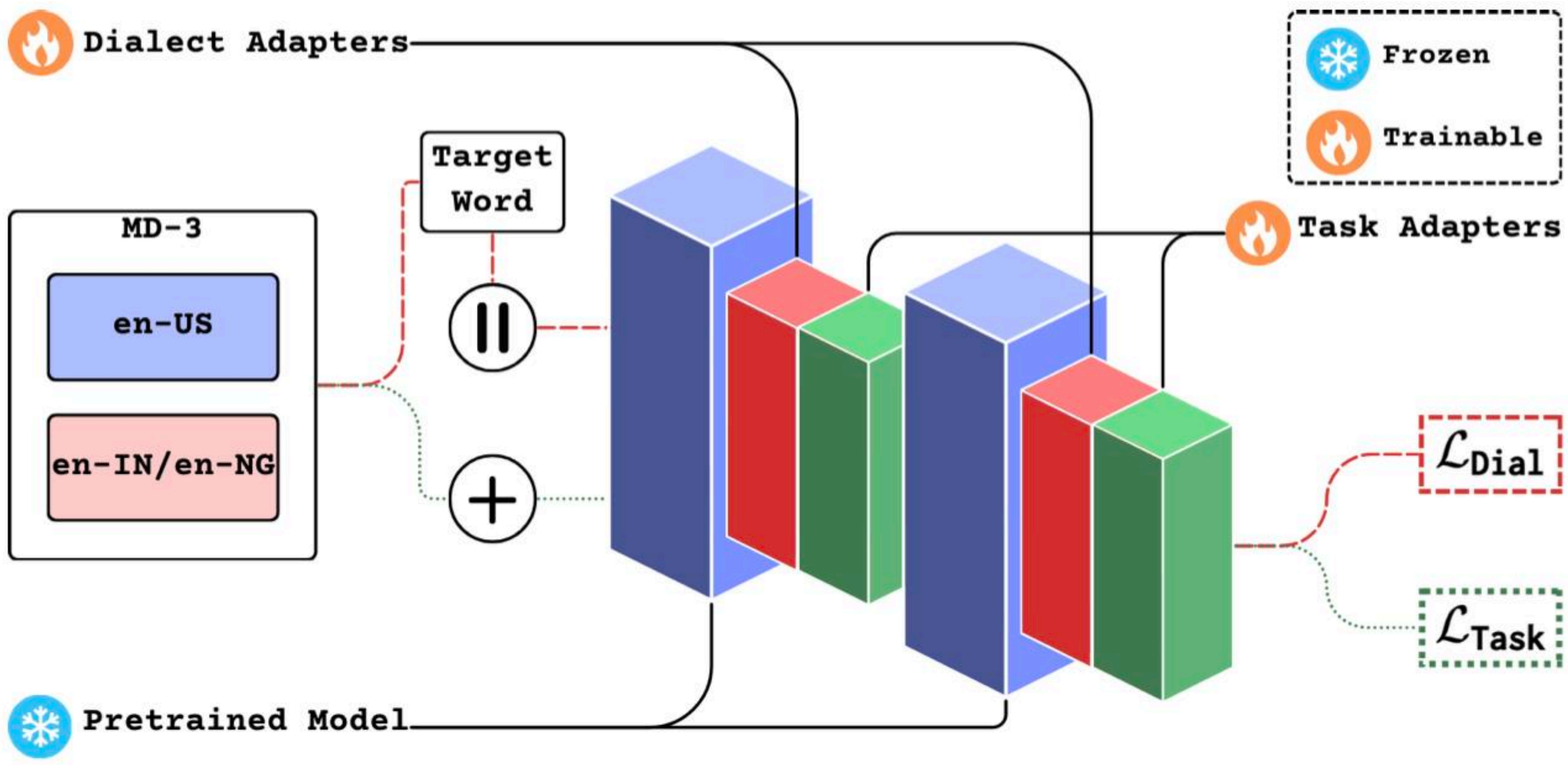
$$\cos\varphi = \frac{\langle [\text{MASK}]_{\text{US}} \cdot [\text{MASK}]_{\text{X}} \rangle}{||[\text{MASK}]_{\text{US}}|| \times ||[\text{MASK}]_{\text{X}}||}$$

Pseudo-parallel Corpus :

- en-US and corresponding en-IN/en-NG conversations for the **same target word**.
- With negative sampling (**different target word**).

Contrastive Loss:

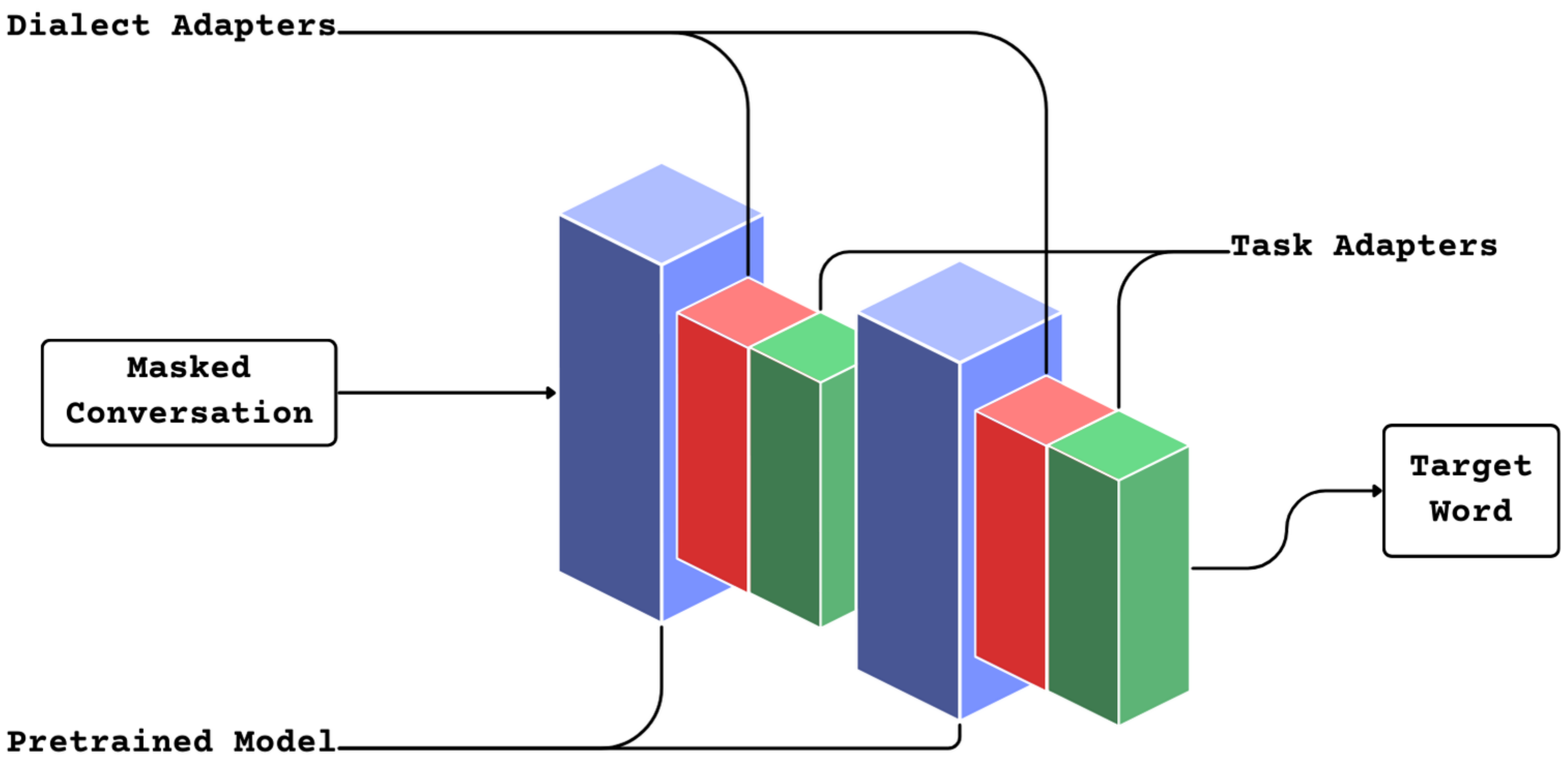
- Frozen representation （$[\text{MASK}]_{\text{US}}$）
- Learnable representation （$[\text{MASK}]_{\text{X}}$）



Low Rank Dialect Robustness for Decoder Models

Inference

Masked Conversation is provided as an input



Task Adapters are stacked on top of Dialect Adapters

LoRDD vs similar work

Approach	Held et al., 2023	Xiao et al., 2023	LoRDD
Models	Encoder-only	Encoder-only	⭐ **Decoder-only**
Dialect Adapter	Invertible Adapters	LoRA adapter	LoRA adapter
Training Method	L2 Loss Critic Network	Hypernetwork over LoRA	⭐ **Cosine Embedding Loss** ⭐ **Instruction Fine-tuning**
Training Data	Synthetic Transformations	Synthetic Transformations	⭐ **Natural conversation pairs**

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.

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.

Experiment Setup

Data

- **en-IN**
- **en-NG**
- **en-US**
- **IN-MV/NG-MV**: Set of transformed en-US conversations using MultiVALUE (**Ziems et al., 2023**).
- **IN-TR** : Set of transformed en-IN conversations using GPT4.

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.

Models

- Mistral 7B Instruct v0.2
- Gemma 2 9B Instruct

mistral
gemma

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. Association for Computational Linguistics.

Skylines and Baselines

Skyline

- Fine-tuned and evaluated on **en-US** conversations.

In-dialect baseline

- Fine-tuned and evaluated on **en-IN/en-NG** conversations

Cross-dialect baseline

- Fine-tuned on:
  - en-US.
  - en-MV.
  - en-TR.
- Evaluated on **en-IN/en-NG** conversations.

LoRDD vs Baslines

LoRDD <u>improves</u> over all <u>baselines</u>.

For en-IN:
- 13.4% on Similarity.
- 28.1% on Accuracy.

For en-NG:
- 11.4% on Similarity.
- 33.8% on Accuracy.

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 TV[illegible] For each mo[illegible]l, 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.

LoRDD <u>closes</u> the gap to the <u>skyline</u>.

For en-IN, the gap is down to:
- 12.0% on Similarity.
- 25.0% on Accuracy.

For en-NG, the gap is down to:
- 5.8% on Similarity.
- 4.5% on Accuracy.

Ablations on LoRDD

Method	Training Data	$\Vert_{\text{Corpus}}$	MISTRAL Similarity	MISTRAL Accuracy	GEMMA Similarity	GEMMA Accuracy	μ Similarity	μ 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).

Synthetic Parallel Corpus

No Dialect Adapter

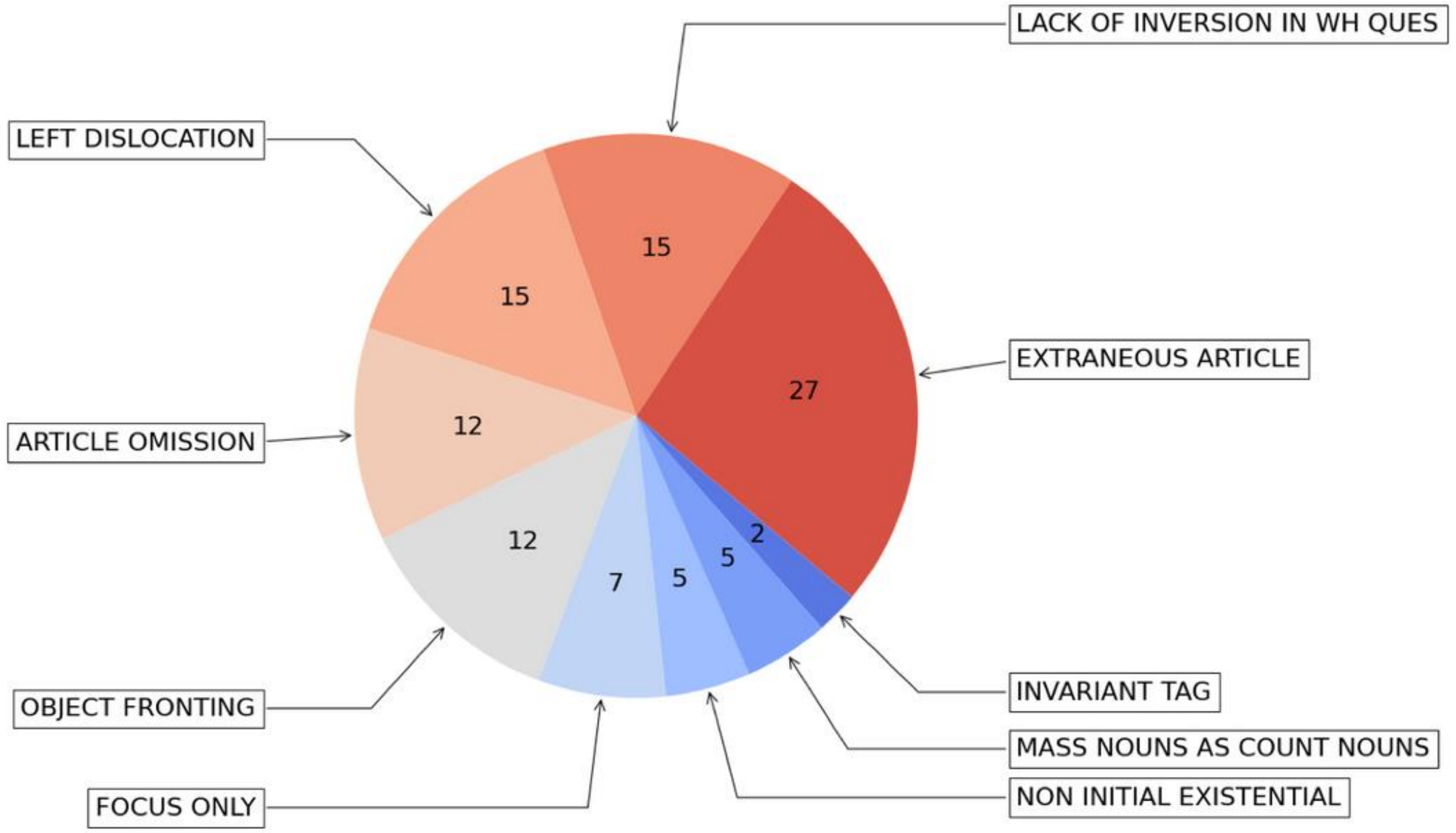
LoRDD, trained using natural conversations, yields the best performance.

All ablations of LoRDD observe a degradation

Particularly lower performances on variants involving synthetic conversations.

Error Analysis

30 en-IN examples misclassified by **gemma**
(trained using LoRDD).



Most common dialect feature is **extraneous articles.**

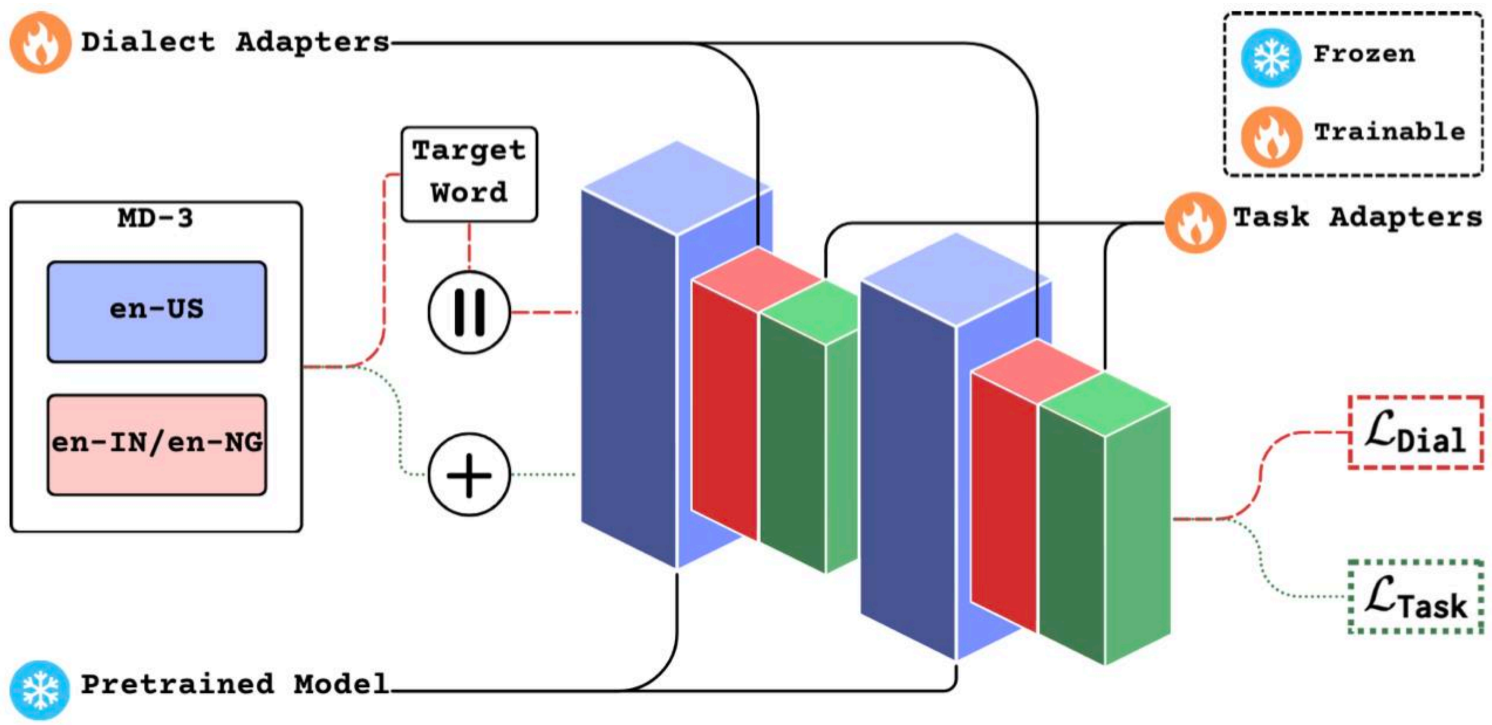
Dialect features as defined by **Demszky et al. (2021)**.

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.

Conclusion

- **Task:** Given a masked dialogue between dialectal speakers playing a game of taboo, predict the target word.
- **Approach (LoRDD):** Dialect adapters for decoder models trained using contrastive loss with pseudo-parallel conversations across dialects.
- **Observations:** LoRDD shows improvement over in-dialect and cross-dialect baselines for en-IN and en-NG.
- **Future work:**
  - Generalisation to other mainstream tasks.
  - Use other forms of parallel corpora.
  - Absence of parallel corpora.

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



Method	Training Data	MISTRAL Similarity	MISTRAL Accuracy	GEMMA Similarity	GEMMA Accuracy	μ Similarity	μ 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)

Preprint

Questions?

Dipankar Srirag | d.srirag@unsw.edu.au

Code