Maastricht University at AMIYA: Adapting LLMs for Dialectal Arabic using Fine-tuning and MBR Decoding

Abdulhai Alali **Abderrahmane Issam**
Department of Advanced Computing Sciences
Maastricht University
{abdulhai.alali@student., abderrahmane.issam@}maastrichtuniversity.nl

Abstract

Large Language Models (LLMs) are becoming increasingly multilingual, supporting hundreds of languages, especially high resource ones. Unfortunately, Dialect variations are still underrepresented due to limited data and linguistic variation. In this work, we adapt a pre-trained LLM to improve dialectal performance. Specifically, we use Low Rank Adaptation (LoRA) fine-tuning on monolingual and English–Dialect parallel data, adapter merging and dialect-aware MBR decoding to improve dialectal fidelity generation and translation. Experiments on Syrian, Moroccan, and Saudi Arabic show that merging and MBR improve dialectal fidelity while preserving semantic accuracy. This combination provides a compact and effective framework for robust dialectal Arabic generation.

1 Introduction

Arabic dialects exhibit substantial variation in vocabulary, morphology, and syntax, making automated generation and translation challenging. Unlike Modern Standard Arabic (MSA), Dialectal Arabic (DA) is underrepresented in NLP resources, leading to difficulties in building models that produce fluent, semantically faithful, and dialectally authentic outputs (Alabdullah et al., 2025). The AMIYA Shared Task (Robinson et al., 2026) targets these challenges by evaluating LLMs on monolingual dialect generation and cross-lingual translation, emphasizing both dialect fidelity and instruction following.

To address these issues, we adapt a Large Language Model (LLM) using parameter-efficient fine-tuning on monolingual and English–Dialect parallel data. We train separate Low-Rank Adaptation (LoRA) adapters (Houlsby et al., 2019; Bapna and Firat, 2019; Hu et al., 2021) for each task (i.e. self supervised training on monolingual data and translation on parallel data), capturing dialectal surface forms and semantic grounding, and combine them using TIES-Merging (Yadav et al., 2023). Additionally, we apply Minimum Bayes Risk (MBR) (Bickel and Doksum, 2007; Kumar and Byrne, 2004; Deguchi et al., 2024) decoding with dialect-aware scoring to select outputs that maximize dialect authenticity during generation.

Our experiments show that merging monolingual and translation-based adapters improves the balance between dialectal fidelity and semantic accuracy. MBR decoding further enhances dialectal authenticity, leading to consistent gains over single-source fine-tuning and standard decoding. While our approach is effective, the following limitations persist: the training data is relatively small, dialect identification metrics may not capture subtle or informal usage, and MBR decoding increases inference time.

2 AMIYA Shared Task

2.1 Task Description

The AMIYA Shared Task focuses on improving LLMs for Dialectal Arabic (DA), which remains significantly underrepresented compared to Modern Standard Arabic (MSA) (Bergman and Diab, 2022). Participants are asked to develop or adapt LLMs that can generate fluent, semantically faithful, and dialectally authentic Arabic across multiple regional varieties.

Systems are evaluated using the AL-QASIDA benchmark (Robinson et al., 2025), which measures dialectal fidelity, generation quality, and robustness to MSA–DA diglossia. Evaluation includes both monolingual dialect generation and cross-lingual settings, such as English–Dialect and MSA–Dialect translation. Performance is assessed using automatic metrics—primarily Arabic Dialect Identification And DIalectnes (ADI2) (Robinson et al., 2025) for dialect fidelity and character-level F-score (chrF++) (Popović, 2015; Popović, 2015)

Proceedings of the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 352–358
March 29, 2026.

for translation quality—as well as human judgments of fluency and instruction adherence.

2.2 Datasets

We participated in the closed data track of the shared task focusing on 3 out of 5 Arabic dialects provided by the task, namely: Syrian, Moroccan and Saudi Arabic. For each dialect, we combine two types of training data: Monolingual dialectal text which consists of unstructured sentences in the target dialect, and Machine Translation (MT) data with English and DA parallel text.

Dialect (Type)	Dataset	# Samples
Syrian (Mono.)	Shami Corpus	25,136
Syrian (MT)	UFAL	120,600
Moroccan (Mono.)	DoDa	10,000
Moroccan (MT)	DoDa	10,000
Saudi (Mono.)	SDC	14,891
Saudi (MT)	SauDial	1,000

Table 1: Datasets used per dialect and supervision type. Shami Corpus (Abu Kwaik et al., 2018), UFAL (Krubiński et al., 2023), DoDa (Darija Open Dataset Contributors, 2023), SDC (TaghreedT/SDC Contributors, 2020), SauDial (Alanazi et al., 2025).

Table 1 summarizes the datasets and sample sizes used for fine-tuning across each dialect and supervision type. While more extensive data is available for most categories, we sub-sample the datasets to maintain computational efficiency and accelerate the experimental process.

2.3 Evaluation Metrics

Evaluation is performed using different metrics depending on the task setting:

Monolingual Dialect Evaluation. For monolingual generation, **ADI2** metric is used. ADI2 score was proposed in Robinson et al. (2025) to measure whether LLMs generate outputs that are dialectal, and whether they are faithful to the specific requested dialect. The level of dialectness is measured using Arabic Level of Dialectness of text (ALDI) (Keleg et al., 2023), and the dialect class C is predicted using a dialect identification baseline model from Nuanced Arabic Dialect Identification (NADI) 2024 shared task (Abdul-Mageed et al., 2024). More formally, ADI2 score (Robinson et al., 2025) is defined as:

$$\text{score}_{\text{ADI2}}(y) = \text{score}_{\text{ALDi}}(y) * \text{score}_{\text{NADI}}(y)_C \quad (1)$$

Cross-Lingual Evaluation. For translation-based and cross-lingual generation tasks, **chrF++** is used for evaluation. chrF++ is well suited for morphologically rich languages such as Arabic, where it captures fine-grained character overlap and is robust to spelling variation, making it appropriate for dialectal evaluation where orthographic inconsistency is common.

3 System Description

3.1 LoRA Fine-tuning

To incorporate dialectal knowledge into the base model, we use parameter-efficient fine-tuning with LoRA. Fine-tuning is performed separately for each dialect and task (i.e. self-supervised and translation), allowing the model to learn different types of information without intervention between them. Table 2 reports our training hyperparamters. These hyperparameters were chosen to ensure stable training under memory constraints while maintaining sufficient capacity for effective dialect adaptation.

Hyperparameter	Value
Max. sequence length	512
Epochs	5
Learning rate	3e-5
Batch size (per device)	2
Effective batch size	32
Precision	BF16

Table 2: Key training hyperparameters.

3.1.1 Monolingual Dialect Fine-tuning

For monolingual adaptation, we fine-tune the model on raw dialectal text without any task-specific prompts. The data consists of standalone sentences written in each dialect, encouraging the model to naturally learn dialect-specific vocabulary, morphology, and sentence structure.

All sentences are tokenized using the JAIS (Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), 2025) tokenizer with a fixed maximum sequence length. We train the model using a standard causal language modeling objective. To make fine-tuning efficient, we apply LoRA adapters (Bapna and Firat, 2019; Houlsby et al., 2019; Hu et al., 2021) to the attention layers of the model and update only these additional parameters during training. Furthermore, we rely on memory-optimization techniques such as gradient accumulation and gradient checkpointing, enabling larger effective batch sizes. The model is trained for multiple epochs using a standard optimization setup.

This approach allows the model to adapt strongly to dialectal surface forms while preserving the general knowledge of the base model.

3.1.2 Translation-Based Fine-tuning

In addition to monolingual data, we fine-tune the model on an English–Dialect parallel dataset. This data exposes the model to aligned semantic content across languages, helping it associate dialectal expressions with their meanings and improving controllability during generation.

We frame translation as an instruction-following task in both directions: English→Dialect and Dialect→English. Each training example includes a natural language instruction specifying the translation direction, followed by the target output. During training, the loss is computed only on the output tokens, while the instruction tokens are masked. This encourages the model to follow instructions without learning to reproduce them.

Tokenization is performed with a fixed maximum sequence length. The same LoRA setup is used as in monolingual fine-tuning to ensure compatibility across training stages. Training is carried out with a more conservative optimization setup than monolingual fine-tuning, focusing on stable learning and semantic alignment rather than aggressive adaptation.

3.1.3 Adapter Merging

The monolingual and translation-based fine-tuning strategies provide complementary supervision. Monolingual fine-tuning emphasizes dialectal fluency and authenticity, while translation fine-tuning reinforces semantic faithfulness and cross-lingual grounding. By training separate LoRA adapters for each dataset, we preserve these distinct signals and later combine them using TIES-Merging. This separation enables fine-grained control over how different sources of supervision contribute to the final dialect-aware model and minimizes intervention between them.

3.2 MBR Decoding with Dialect-Aware Scoring

While fine-tuning and merging improve the model's internal dialect representations, decoding decisions still play a crucial role in output quality. We therefore apply MBR decoding using the *mbrs*[1] library to explicitly optimize for dialectness at inference time.

For each input prompt, the model generates a set of $N = 20$ candidate responses via stochastic sampling, then each candidate is scored independently using the ADI2 metric. Finally, the candidate with the highest score is selected as the final output.

3.3 Adapter Merging and MBR

TIES-based adapter merging integrates complementary dataset supervision at the parameter level, producing a compact yet expressive dialect-aware model. MBR decoding complements this by enforcing dialectal fidelity at generation time, explicitly selecting outputs that maximize dialectness. Together, fine-tuning, TIES-Merging, and MBR decoding form a unified framework that yields more consistent and authentic dialectal generation than any single technique in isolation.

4 Results

The experiments were conducted exclusively on Syrian and Moroccan DA and subsequently applied to the remaining dialects for the final submission, which was trained separately per dialect. This section presents a detailed evaluation of our dialect-aware generation framework. We report results across model variants, data configurations, and decoding strategies, with the goal of understanding (1) the impact of model choice, (2) the role of different supervision signals, and (3) the effectiveness of adapter merging and MBR decoding. The evaluation datasets used are the default datasets provided by AL-QASIDA [2].

4.1 JAIS-2 vs. LLaMA 3.2

We begin by comparing two LLMs, JAIS-2[3] (Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), 2025) and LLaMA 3.2[4] (Grattafiori et al., 2024), to determine the most suitable backbone for Arabic dialect generation. For a fair comparison, both models are fine-tuned using the same data configuration: a merged setup that combines monolingual dialect data with English–Dialect parallel (MT) supervision. In addition, decoding is performed using MBR decoding with ADI2 score.

Models are evaluated using ADI2 for monolingual dialect generation and chrF++ for trans-

[1] https://github.com/naist-nlp/mbrs

[2] https://github.com/JHU-CLSP/al-qasida

[3] https://huggingface.co/inceptionai/Jais-2-8B-Chat

[4] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

lation. Table 3 presents the results. On monolingual dialect generation, LLaMA 3.2 achieves a substantially higher ADI2 score (i.e. 0.78), indicating strong dialectal surface realization and fluency. However, its performance drops sharply in translation, with a significantly low chrF++ score (i.e. 0.14), suggesting weak semantic alignment when translating from English into dialectal Arabic. In contrast, JAIS-2 exhibits a more balanced performance. While its ADI2 score (0.33) is considerably lower than that of LLaMA 3.2 for monolingual generation, JAIS-2 achieves a much higher chrF++ score (0.43) on MT-based generation. This indicates stronger semantic fidelity and better handling of translation supervision.

Model	ADI2	chrF++
LLaMA 3.2	**0.78**	0.14
JAIS-2	0.33	**0.43**

Table 3: Comparison between JAIS-2 and LLaMA 3.2 after fine-tuning, TIES-Merging and generation with MBR decoding on **Syrian DA**

.

On overall, although LLaMA 3.2 excels in dialectal surface form generation, its poor cross-lingual performance limits its usefulness for translation-driven dialect generation. Given our goal of building a dialect-aware system that remains reliable across both monolingual and cross-lingual scenarios, we select JAIS-2 as the backbone for all subsequent experiments.

4.2 Effect of Fine-tuning Data and Adapter Merging

In this section, we analyze the impact of different fine-tuning strategies on JAIS-2. We report the results of JAIS-2 base model, JAIS-2 fine-tuned on either monolingual or translation data, and JAIS-2 with TIES-Merging. The results in Table 4 show that monolingual fine-tuning substantially improves ADI2 scores, indicating stronger dialectal identity and linguistic conformity. Furthermore, fine-tuning on parallel translations significantly improves chrF++ scores, reflecting improved semantic faithfulness and cross-lingual grounding. More importantly, merging monolingual and MT models using TIES-Merging consistently improves the balance between dialectal authenticity and semantic accuracy, leading to the best chrF++ and the second best ADI2 score performance.

Configuration	ADI2	chrF++
JAIS-2 (Base)	0.18	0.31
+ Monolingual FT	**0.44**	0.33
+ MT FT	0.26	0.42
+ TIES-Merging	0.38	**0.44**

Table 4: Effect of monolingual and MT task fine-tuning and TIES-Merging on **Moroccan DA**. Merging both tasks (i.e. TIES-Merging) leads to the best performance on chrF++ and the second best on ADI2 score.

4.3 MBR Decoding with Dialect-Aware Objectives

While adapter merging improves the model's internal representations, standard decoding does not always select the most dialectally appropriate output. To address this, we apply MBR decoding with different objectives. MBR requires a target metric to score the candidate generations. We experiment with using ADI2 to improve dialectal fidelity, chrF++ to improve cross-lingual grounding, and their combination. As shown in 5, MBR decoding with ADI2 achieves the best overall balance, improving monolingual ADI2 to approximately 0.51 while also increasing MT ADI2 to 0.36. This represents a substantial improvement over standard decoding and demonstrates that dialect-aware reranking can recover dialectal authenticity without sacrificing semantic grounding. In contrast, chrF++-optimized MBR and the combined objective favor translation quality: they achieve higher chrF++ scores (0.42 and 0.41, respectively) but lead to a significant drop in monolingual ADI2 (0.24 and 0.29). These results indicate that chrF++-centric objectives bias the model toward more neutral or standardized Arabic forms, reducing dialectal distinctiveness. Based on these findings, and given our emphasis on dialect fidelity while maintaining acceptable translation performance, we select ADI2-based MBR decoding for the final submission.

Decoding Strategy	Mono ADI2	MT ADI2	chrF++
Standard decoding	0.38	0.27	**0.44**
MBR (ADI2)	**0.51**	0.36	0.40
MBR (chrF++)	0.24	0.30	0.42
MBR (ADI2 + chrF++)	0.29	**0.37**	0.41

Table 5: Effect of MBR decoding objectives on JAIS-2 after fine-tuning and TIES-MErging on **Moroccan DA**. Using ADI2 as an objective strikes the best balance in ADI2 and chrF++ performance.

4.4 Final Submission

Across all experiments, the best-performing configuration is: JAIS-2 independent fine-tuning on monolingual and translation data, merging with TIES technique, and decoding using MBR with ADI2 as an objective metric. This configuration achieves the strongest balance between dialectal authenticity and cross-lingual grounding, outperforming all alternatives in terms of combined monolingual and MT performance. Consequently, our *primary* submission applies this methodology individually to the Moroccan, Syrian, and Saudi dialects. While joint training across all dialects may offer further gains, we defer this exploration to future work.

5 AMIYA Shared Task Official Results

Table 6 summarizes the official automatic evaluation results of our *primary* submission. The results demonstrate strong generalization across dialects and task settings, confirming that the combination of parameter-efficient fine-tuning, TIES-Merging, and dialect-aware MBR decoding provides a robust and effective solution for DA generation.

Dialect	ADI2	DA→EN	EN→DA	DA→MSA	MSA→DA
Moroccan	**0.679**	49.93	30.02	39.53	33.77
Syrian	0.389	**51.89**	**34.44**	**43.42**	**40.33**
Saudi	0.464	0.03	19.82	37.21	24.23

Table 6: Automatic evaluation results using ADI2 and chrF++.

Besides automatic evaluation, our model generations were human evaluated for fluency and adherence to DA instructions. Table 7 shows that our highest human evaluation performance is on the Moroccan dialect.

Dialect	Adequacy	Fluency
Moroccan	**1.97**	**3.37**
Syrian	1.146	2.625
Saudi	1.122	2.378

Table 7: Human evaluation scores.

Across submissions from other teams, our system achieved the highest ADI2 for Syrian and Saudi, and the highest chrF++ scores on translation from English and MSA into Syrian (ENG→DA and MSA→DA). On Moroccan Arabic, our model performs best on the human evaluation of fluency.

6 Conclusion

We presented a method for improving Dialectal Arabic generation by combining fine-tuning on dialectal and translation data, LoRA adapter merging, and MBR decoding. This approach helps the model produce outputs that are both fluent in the target dialect and faithful to the input meaning. Our experiments across three dialects show that this combination works better than using any single technique on its own, providing a practical way to make LLMs more dialect-aware.

Limitations

This work has the following limitations: Our decoding strategy depends on automatic dialect identification (ADI2) scores, which may not always capture subtle or informal dialectal usage. The training data is limited in size and may not cover all linguistic variation within each dialect, especially code-switching and colloquial expressions. Finally, both ADI2 computation and minimum Bayes risk (MBR) decoding increase inference time. ADI2 requires an additional forward pass, and the cost of MBR grows with the number of candidate hypotheses generated per input, resulting in a several-fold slowdown compared to standard decoding. As a result, the approach may be less practical for real-time or low-latency applications.

References

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. NADI 2024: The fifth nuanced arabic dialect identification shared task. In *Proceedings of The Second Arabic Natural Language Processing Conference (Arabic-NLP 2024)*, pages 709–728, Bangkok, Thailand. Association for Computational Linguistics.

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. Shami: A corpus of levantine arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA). Dataset includes Levantine Arabic tweets covering Palestinian, Jordanian, Syrian, and Lebanese dialects.

Abdullah Alabdullah, Lifeng Han, and Chenghua Lin. 2025. Advancing dialectal arabic to modern standard arabic machine translation. *arXiv preprint arXiv:2507.20301*. Dialectal Arabic–MSA MT challenges and resource-efficient strategies.

Naif Alanazi, Mohammed Al-Batineh, and Hussein Abu-Rayyash. 2025. Saudial: Saudi arabic dialects game localization dataset. https://data.mendeley.com/datasets/mzdwkb2t6d/2. Parallel Saudi dialect text dataset (English, MSA, Saudi varieties) with cultural context, age ratings, and dialect notes. CC-BY 4.0 license.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. 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 2414–2425.

A. Bergman and Mona Diab. 2022. Towards responsible natural language annotation for the varieties of Arabic. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 364–371, Dublin, Ireland. Association for Computational Linguistics.

Peter Bickel and Kjell Doksum. 2007. *Mathematical Statistics: Basic Ideas and Selected Topics.*, volume 56.

Darija Open Dataset Contributors. 2023. Doda: Darija open dataset. https://github.com/darija-open-dataset/dataset/tree/main?tab=readme-ov-file#citation. Dataset of 48,849 Darija sentences, with English translations. CC-BY-NC license. Downloadable via link. High quality.

Hiroyuki Deguchi, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. mbrs: A library for minimum Bayes risk decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 351–362, Miami, Florida, USA. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. *Preprint*, arXiv:1902.00751.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. https://arxiv.org/abs/2106.09685. Presented at ICLR 2022; parameter-efficient adaptation method for large pretrained models.

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. ALDi: Quantifying the Arabic level of dialectness of text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10597–10611, Singapore. Association for Computational Linguistics.

Mateusz Krubiński, Hashem Sellat, Shadi Saleh, Adam Pospísil, Petr Zemánek, and Pavel Pecina. 2023. Multi-parallel corpus of north levantine arabic. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, pages 411–417. Association for Computational Linguistics. Corpus includes 120,600 multiparallel sentences in English, French, German, Greek, Spanish and MSA manually translated into North Levantine Arabic.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Mohamed bin Zayed University of Artificial Intelligence (MBZUAI). 2025. Inception, cerebras and mbzuai release Jais 2 – the next generation arabic open-weight llm. https://mbzuai.ac.ae. Accessed: 2026-01-21.

Maja Popović. 2015. chrf: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2015. chrf: Character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Nathaniel Robinson, Shahd Abdelmoneim, Anjali Kantharuban, Otba Alsboul, Salima Lamsiyah, Kelly Marchisio, and Kenton Murray. 2026. AMIYA shared task: Arabic Modeling In Your Accent at VarDial 2026. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, Rabat, Morocco. Association for Computational Linguistics.

Nathaniel R. Robinson, Shahd Abdelmoneim, Kelly Marchisio, and Sebastian Ruder. 2025. Al-qasida: Analyzing llm quality and accuracy systematically in dialectal arabic. *Preprint*, arXiv:2412.04193.

TaghreedT/SDC Contributors. 2020. Saudi dialect corpus (sdc). https://github.com/TaghreedT/SDC/blob/main/SDC.txt. A 210,396-word corpus of Saudi Arabic social media posts collected from platforms such as Facebook and Twitter. High quality text format, accessible via link; please cite Tarmom et al. (2020) when using this corpus.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. TIES-merging: Resolving interference when merging models. In *Thirty-seventh Conference on Neural Information Processing Systems*.