NUS-IDS at AMIYA/VarDial 2026: Improving Arabic Dialectness in LLMs with Reinforcement Learning

Sujatha Das Gollapalli,[1] **Mouad Hakam,**[1] **Mingzhe Du,**[1,2] **See-Kiong Ng**[1]
[1]Institute of Data Science, National University of Singapore
[2]College of Computing and Data Science, Nanyang Technological University
{idssdg,mouad.hk,mingzhe,seekiong}@nus.edu.sg

Abstract

In this paper, we describe models developed by our team, NUS-IDS, for the *Closed Data* track at the Arabic Modeling In Your Accent (AMIYA) shared task at VarDial 2026. The core idea behind our solution involves data augmentation enabled by a dialect classifier trained on AMIYA data. We effectively combine various translation, summarization, and question answering prompts with the training data to form dialectal prompts for use with state-of-the-art Large Language Models (LLMs). Next, dialect predictions on outputs from these LLMs are used to compile preference data for Reinforcement Learning (RL). We report model performance on dialectal Arabic from Egypt, Morocco, Palestine, Saudi Arabia, and Syria using FLORES+, a benchmark dataset for multilingual machine translation. Our experiments illustrate that though our RL models show significant performance gains on dialectness scores, they underperform on translation metrics compared to base LLMs.

1 Introduction

Recent AI innovation in form of Large language models (LLMs) has shown revolutionary potential for solving various data processing and language understanding tasks. Indeed, state-of-the-art (SOTA) proprietary LLMs from OpenAI, Google, etc. demonstrate competitive performance on various linguistic tasks including translation, summarization, question answering, and sentiment analysis even in zero-shot settings using simple prompts (Vatsal and Dubey, 2024). However, ongoing multilingual research studies note that SOTA LLMs are not uniformly proficient in all languages and their error rates are often significantly higher when handling instructions in non-Latin scripts such as Arabic, Russian, and Hindi (Hasan et al., 2024; He et al., 2024). Similarly, there is room for improvement in current LLM capabilities for handling "dialects" or *variants of a language that are characteristic of a particular group of the language's speakers* as shown in lower NLP task performances when dialects are involved (Joshi et al., 2025).

The lower performance of SOTA LLMs for non-English languages, and dialect variants has often been attributed to the absence of large-scale resources for these language variants available for LLM pretraining (Inoue et al., 2021). Indeed, for languages such as Arabic, though numerous region-specific dialects are used in spoken and informal communications (such as tweets and blogs), most available corpora are based on formal settings (such as news and education) for which Modern Standard Arabic (MSA) is employed. Possibly as a result of this disparity between the scale of MSA and Dialectal Arabic (DA) corpora available for LLM pretraining, a recent study noted that SOTA LLMs do not naturally respond in dialectal Arabic and instead employ MSA even when the prompts used are in DA (Robinson et al., 2025). Motivated by these observations on LLMs for Dialectal Arabic and to encourage research on handling the same, the Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial) 2026 included the Arabic Modeling In Your Accent (AMIYA) Shared Task (Robinson et al., 2026).

Task Description and Datasets: We (the NUS-IDS team) participated in the AMIYA Shared Task, which requires participants to contribute LLMs trained or adapted for Dialectal Arabic. Specifically, in the *Closed Data* track, task participants are required to only use the provided training data to fine-tune open-source LLMs such as those from the Llama[1] and Qwen families.[2]

The training data released for this competition includes sentences in Dialectal Arabic from five countries: Egypt, Morocco, Palestine, Saudi

[1]https://huggingface.co/meta-llama
[2]https://huggingface.co/Qwen

Proceedings of the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 365–372

Arabic and Syria (*Egy/Mar/Pse/Sau/Syr*) and includes sentences from parallel corpora (same sentences written in multiple dialects) such as MADAR (Bouamor et al., 2018). In addition, the development split from the FLORES+ benchmark (Perez-Ortiz et al., 2024) was made available for model tuning with final evaluation centered around a different set of dialectal prompts from the benchmark.

The metrics proposed by task organizers for evaluation include the ADI2 dialect fidelity score (Robinson et al., 2025), ChrF++ scores (Popović, 2017) used in machine translation (MT) as well as human evaluation of fluency and adherence to DA instructions. Note that the translation score ChrF++ involves the computation of character n-gram overlap with ground-truth (available for MT data) whereas ADI2 combines dialectness (MSA is considered non-dialectal) with whether the output was in the specific regional dialect (requested in the prompt) to assign a "reference-free" dialect fidelity score. For the experiments in this paper, we directly employ the implementations provided by the task organizers in AL-QASIDA[3] for computing the above metrics. We refer the reader to the Task Overview Paper for details on the datasets and metrics used for this task (Robinson et al., 2026).

2 Data Augmentation

Building on the observations from earlier studies (Robinson et al., 2025) and our own investigations with smaller open-source LLMs, we note the tendency of LLMs to respond in Modern Standard Arabic (MSA) instead of Dialectal Arabic (DA). We posit that this limitation may be addressed by incentivizing the LLM to respond in DA when the instruction/prompt calls for it. For example, when the prompt is in DA (without an explicit dialect mentioned) or prompts requesting the LLM output to be in a specific dialect. That is, a pairwise preference dataset that can be used to train a given LLM with Reinforcement Learning (RL) to prefer a DA response over an MSA one is required for suitably aligning the LLM. Indeed, this notion of fine-tuning LLMs using preference data compiled from human feedback through reinforcement learning (RLHF) is a widely-used technique for aligning LLM behavior with user intent in various scenarios (Ouyang et al., 2022; Lee et al., 2024). *How can we build preference pairs data for RL using the given AMIYA datasets?* We employ the data augmentation pipeline illustrated in Figure 1 for this precise purpose.

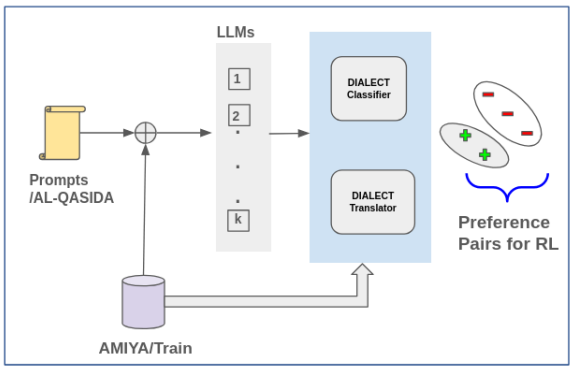

Figure 1: Data Augmentation Pipeline

First, the training data from AMIYA was used to learn dialect classification and translation models by instruction tuning open-source LLMs from Meta[4] and Qwen.[5] We employed these open-source LLMs and proprietary models (GPT-4o[6] and Gemini-2.5-Flash[7]) for creating preference pairs as follows:

1. The summarization, paraphrasing, and story writing prompts from the AL-QASIDA dataset were combined with the sentences from the AMIYA training data sampled after applying thresholds on dialectness scores (Keleg et al., 2023) and number of words.[8]

2. The (prompt, input) pairs compiled in the above step were used with our selection of LLMs to obtain multiple output responses for the same pair.

3. The dialect classifier is used to label the dialects of LLM responses from the second step. If the label does not match the desired dialect, we employ the translation model to convert the response to the desired dialect and re-check

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

[4] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[5] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[6] https://platform.openai.com/docs/models/gpt-4o

[7] https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash

[8] For example, short sentences (with fewer words) can be used to complete a story whereas long texts comprise inputs to summarization prompts.

using our classifier. In this fashion, we obtain "positive" and "negative" responses for the same LLM input.

4. Finally, positive and negative responses obtained above are used to form preference pairs for RL. That is, if d_i is the desired dialect, and R refers to an LLM response, $(prompt, input, R_{d_i}) \succ (prompt, input, R_{d_j})$ where $j \neq i$.

3 Experiments and Results

Datasets and Setup: Our dialect classifier was trained on $\sim$ 300K sentences obtained by randomly sampling around $\sim 50K$ sentences for each dialect from the training data in AMIYA. For training our dialect translation model, we used $\sim$ 697K sentences compiled from the parallel data provided in the training data and employing the NLLB translation model (Costa-jussa et al., 2024) when parallel translations were unavailable.[9] In particular, we used NLLB for translating from MSA to English or vice-versa when only the MSA-DA pair or the English-DA pair was included for a given sentence in the training data from AMIYA.

For compiling preference pairs, we used about 24 prompts available for each dialect in AL-QASIDA[10] and randomly sampled about 1000 sentences with dialectness scores (Keleg et al., 2023) above 0.36 in the training data, for each dialect. Additionally, we employed translation prompts for MSA to DA using random samples of sentences provided in the training data. Overall our data augmentation pipeline resulted in a set of approximately 9.6K preference pairs for use with RL. The datasets compiled by us for the AMIYA competition along with descriptions of the process are shared to enable further research on the topic.[11]

We used the LoRA (low-rank adaptation) training scripts provided in LlamaFactory (Hu et al., 2022; Yao et al., 2024) for all our supervised fine-tuning and RL experiments. The number of epochs were set to three and default values provided in the training scripts in LlamaFactory were used for other parameters. For the rest of this paper, we refer to the base models, Llama-3.1-8B-Instruct[4] and Qwen-2.5-7B-Instruct,[5] as **llama** and **qwen**, respectively. The base models fine-tuned with translation data are referenced as **llama-T** and **qwen-T**, respectively. We tested RL using the base models as well as our fine-tuned translation models as starting points. Models after Reinforcement Learning using the direct preference optimization algorithm (Rafailov et al., 2023) are indicated using the **+RL** suffix. All our experiments were performed on an NVIDIA A6000 server with 2 GPUs each having 48GB RAM .

Classification Performance: We show the F1 scores on a sample held-out test dataset from AMIYA that contains about 200 sentences for each dialect *Egy/Mar/Pse/Sau/Syr* as well as MSA in Figure 3. Our dialect classifier which is a Llama 3.1-8B-Instruct model fine-tuned (**FT-Llama**) on the considerably large sentence-level training data from AMIYA yields high F1 scores with the performance on the Saudi Arabian dialect being the lowest at 0.83 F1 and the Palestinian dialect being the highest at 0.94 F1. Our fine-tuned dialect classifier is able to accurately discriminate among the five dialects and MSA, and shows remarkable performance gains compared to the zero-shot setting with the base Llama model (FT-Llama versus Llama/zero in Figure 3).

We tested the NADI baseline[12] employed in previous works for dialect prediction (Abdul-Mageed et al., 2024; Robinson et al., 2025). Our **FT-Llama** predictor is instruction fine-tuned using a multiple-choice QA-style instruction "Which of the following dialects (a) Egyptian (b) Palestianian...is most applicable...". We specifically trained our model for the five dialects and MSA targeted in the AMIYA competition using the dataset shared by them. In contrast, the NADI baseline comprises a fine-tuned MARBERT model (Abdul-Mageed et al., 2021) for discriminating up to 18 dialects of Arabic (and does not include MSA). As such, on our evaluation dataset, the F1 scores with NADI were 0.68, 0.80, 0.59, 0.79, 0.83 for Egyptian, Moroccan, Palestinian, Saudi, Syrian dialects, respectively and are significantly lower than those obtained with our **FT-Llama** model on these dialects.

Translation Performance: We evaluated our models on the FLORES+ dataset (Perez-Ortiz et al., 2024), using the dev portion included in

[9] https://huggingface.co/facebook/nllb-200-3.3B

[10] https://github.com/JHU-CLSP/al-qasida/blob/main/data_processing/prompt_templates/json/dial8.json

[11] https://github.com/mouad157/Amiya26

[12] https://huggingface.co/AMR-KELEG/NADI2024-baseline

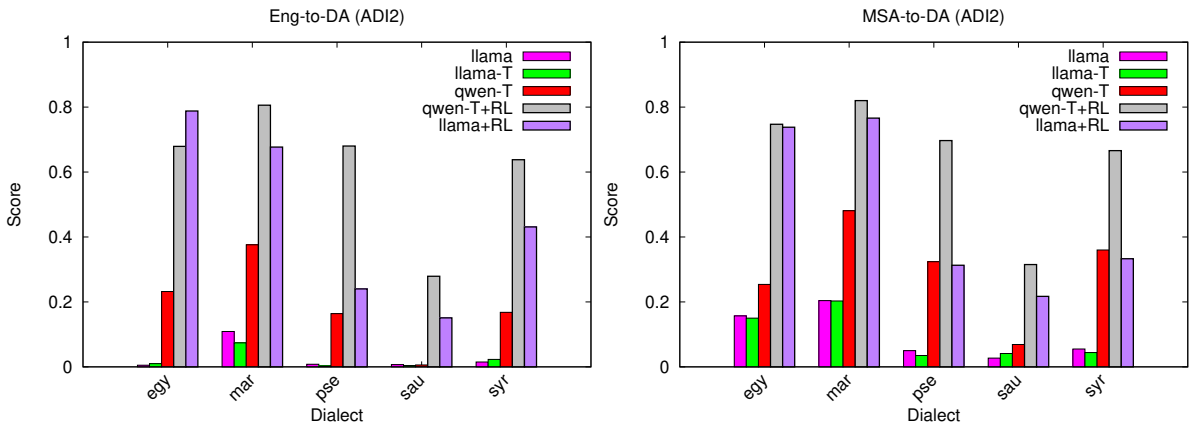

Figure 2: ADI2 Macro Scores for Eng-DA and MSA-DA in FLORES/dev

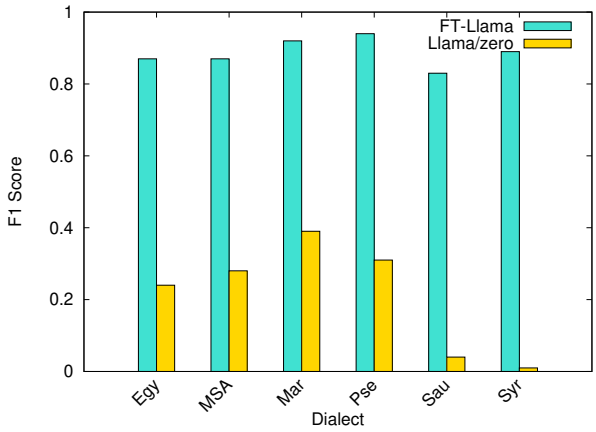

Figure 3: Dialect Classification Performance

AL-QASIDA and provided in the competition. The ADI2 macro scores for the different dialects are highlighted in Figure 2. As observed in previous studies, the dialectness scores with the base models (**llama** and **qwen**) are very low highlighting the tendency of these LLMs to "avoid" dialectal Arabic in their responses. Fine-tuning the base models with translation data yields considerable improvements for most dialects in **qwen-T** but not in **llama-T**. To test if this difference is a result of error-prone augmented data from NLLB model[9] used for training the translation variants of the base models, we tested the models using only the "clean" translation data for which ground truth was available without employing NLLB for augmentation.

The dialectness macro scores for the different MT directions (FLORES data) are shown for the Qwen and Llama base models as well those trained with augmented ('-Aug') and non-augmented translation data ('-Clean') in Figure 7. As can be seen in this figure, for Qwen models, fine-tuning with translation data improves performance on most dialects, with NLLB augmented data producing significant increases in dialectness scores. However, fine-tuning with translation data did not improve performance when the base model is from Llama. We hope to investigate these contradictory results in Llama versus Qwen in our on-going research beyond this paper. For the present, we speculate if this lack of improvement in Llama is possibly due to having seen the data during model pretraining and potential overlap with NLLB datasets (since NLLB is also a model released by Facebook/Meta).

In Figure 2, we note from the Eng-DA and MSA-DA plots that ADI2 macro scores significantly improve with RL. However, the overall dialectness scores are lowest for the Saudi Arabian dialect. The character n-gram overlap scores (ChrF++) (Popović, 2017) used in MT evaluation are shown for MT directions *Eng-DA, MSA-DA, DA-Eng, DA-MSA* in Figures 4 and 5, respectively. In these plots we note that the MT performance notably degrades for all dialects after RL. This reduction in scores reveals a core weakness in our data augmentation pipeline–focusing on using "dialect" accuracy to form preference pairs for RL is inadequate. With dialect correctness alone as an incentive, the models seem to veer off their objective to optimize the quality of the generated response. We confirmed this possibility by manually analyzing a small sample of model outputs for each MT direction.

In Figure 6, we show sample generated responses from our **qwen-T+RL** and **llama+RL** models alongside the ground-truth references. While the translations are reasonably good in some cases (Set 2), our models seem to make errors when

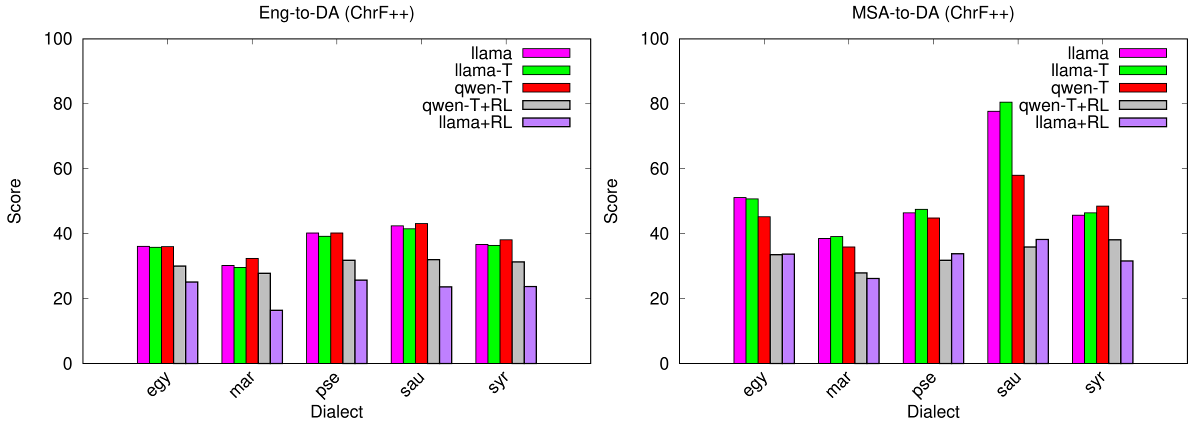

Figure 4: ChrF++ scores for Eng-DA and MSA-DA in FLORES/dev. llama refers to the base model from llama[4], whereas llama-T and qwen-T are the models fine-tuned with translation data and '+RL' indicates models after reinforcement learning with dialect preference pairs.

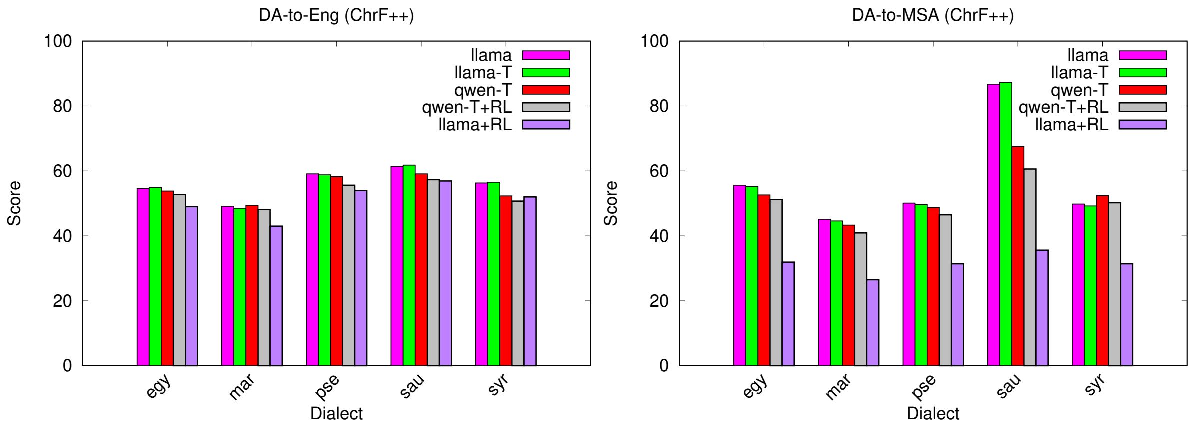

Figure 5: ChrF++ scores for DA-Eng and DA-MSA in FLORES/dev.

(Set 1)
msa-sau
Reference= ما لم تكن دبلوماسيًا، فإن العمل في الخارج يعني أنه سيتعين عليك بشكلٍ عامٍ تقديم ضريبةِ دخلٍ في البلد الذي تُقيم فيه.
Generated= وش درس هذي؟ اللي مو دبلوماسي، يروح يشتغل برا يعني...

eng-mar
Reference= من بين التطبيقات اللي مجهدين فهاد الفئة هو كوكل ترانسليت اللي تيوفر الترجمة بدون الاتصال بالأنترنت من بعد متيتيليشارجي المعطيات اللغوية.
Generated= واحد من أحسن الأبليكيات فهاد لقسم هيا ترجمان ديال غولغل، اللي كاتخلّي الترجمة فالعطلة مورا ماتندمش المعلومات ديال اللهدرا لي بغيتي.

egy-eng
Reference= Using GPS navigation apps on your smartphone can be the easiest and most convenient way to navigate when out of your home country.
Generated= Usam',b taktilemat nassem system JBES ely a3zy hawfik eldhaky, mosknob yoferlek tareefa sahla we mreha3 ely a3shan eltenkob fe beld mo' beldek.

(Set 2)
pse-msa
Reference= احتوت لونو على من 120 إلى 160 متراً مكعّباً من الوقود على متنها، عند تحطمها ودفع الرياح الشديدة والأمواج العالية لها لتصطدم بحائل الأمواج.
Generated= إذا كان فيه حوالي عشرة عشرة أو عشرين متر مكعب من البنزين كان رح يقدر يواجه الرياح العالية والبحر العالي وبعدين رح يروح بسرعة

syr-eng
Reference= Evidence thereafter indicated test papers were tampered with Hall, along with 34 other education officials, was indicted in 2013.
Generated= It turned out that Holt had been cheating on his tests with 34 other nonlicensed educators in 2013.

Figure 6: Example generated responses from RL-based models. Except the third (egy-eng) example that was generated by the *llama+RL* model, the rest are the outputs from our *qwen-T+RL* model

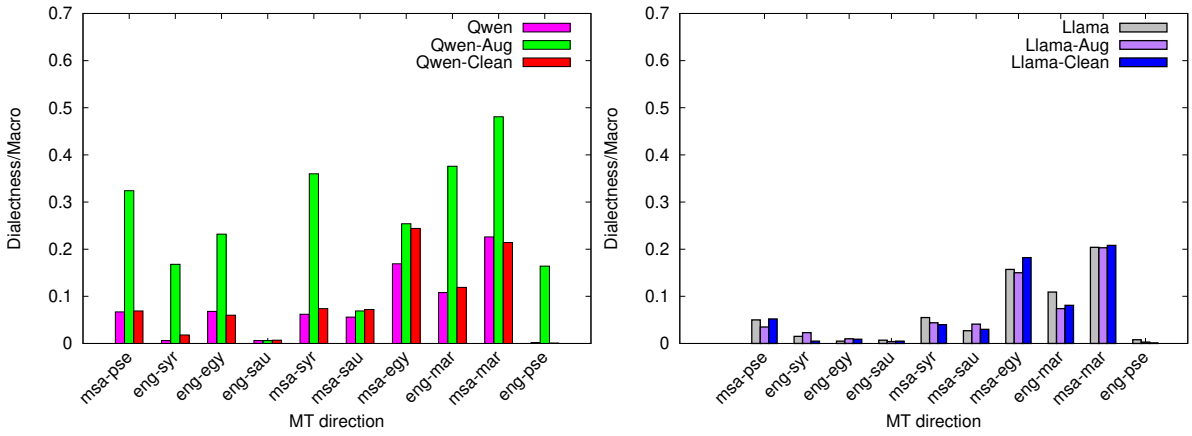

Figure 7: Dialectness Macro scores on FLORES/dev are shown for our translation models.

Model	SpBLEU	chrF++	AvgALDi
llama	15.48	32.09	0.1659
qwen	10.12	**33.35**	0.1429
llama-T	16.67	32.88	0.1663
qwen-T	**16.71**	30.22	0.2441
⋆*qwen3-T*	12.43	26.52	0.2194
⋆*llama+RL*	7.74	23.03	**0.6949**
qwen+RL	7.75	26.54	0.4192
llama-T+RL	8.07	24.93	0.6491
⋆*qwen-T+RL*	10.30	23.89	0.5780

Table 1: Validation Performance across all models

the translation inputs involve numeric information such as times and numbers and proper names. For several cases, (in models with and without RL), characters are indefinitely repeated resulting in garbage responses ('msa-sau' example in Set 1). RL models seem to have a propensity to choose dialectal responses. This aspect was noticed while testing DA-Eng translation data where several LLM outputs are in dialects that can be written using the Latin script ('egy-eng' example in Set 1) instead of in English as required in the prompt.

We note from Figures 4 and 5 that the translation scores on the Saudi dialect for base models for MT directions (MSA-DA and DA-MSA) are unusually higher than for the rest of the dialects. This peculiarity coupled with the lower classification performance in Figure 3 as well as the low dialectness scores in Figure 2 for this dialect is indicative of the proximity between the Saudi dialect and MSA compared to the other dialects considered in this evaluation (Alsudais et al., 2022) and also seems to be an artifact of the FLORES dataset which mostly contains MSA for Saudi portion of the dataset (Perez-Ortiz et al., 2024).

Model selection for the competition: For selecting the top-3 models for consideration in the final evaluation of the shared task, we set aside about 50 examples for each dialect from our data augmentation pipeline for validation, by treating the GPT-generated response in the desired dialect (as identified by our classifier) as "gold/correct" response. All model variations were tested on this subset and the models indicated with an asterisk (⋆) in Table 1 were chosen for final evaluation. **llama+RL** outputs correspond to our *primary* run in AMIYA whereas '**qwen-T+RL**' and '**qwen3-T**' outputs correspond to the *contrastive1* and *contrastive2* runs, respectively.[13]

Our test/competition performance (Robinson et al., 2026), provided by the task organizers, is shown in Table 2. In this table, we included the best scores among our different models for each metric and the overall best scores from the competition. The primary and contrastive runs are indicated using 'p' and 'c', respectively. Mirroring our own evaluation on the FLORES dataset, our primary model obtains a significantly high score on the ADI2 metric but the translation scores are significantly lower than the best scores in the competition.

4 Conclusions

We participated in the Dialectal Arabic modeling shared task (AMIYA@VarDial 2026) in the *Closed Data* track. To handle current LLM models' tendency to respond in Modern Standard Arabic

[13]The 'qwen3-T' model is similar to qwen-T but uses the base model from https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507.

Team	Overall ADI2	DA-ENG	ENG-DA	DA-MSA	MSA-DA
NUS-IDS	**0.629** (p)	21.338 (c1)	15.764 (c1)	19.738 (c1)	17.280 (p)
MBZUAI	0.452 (c2)	**53.436** (c1)	**34.314** (c1)	**50.639** (c1)	**43.728** (c1)

Table 2: Our best performing scores are compared with the overall best scores (in **bold** font) from the competition. Columns 3-6 show ChrF++ scores for the MT directions indicated in the header

rather than dialectal variants despite prompting, we adopted a Reinforcement Learning based solution. Our main contribution to enable RL pertains to the development of a novel data augmentation pipeline that uses the training data from the competition to learn a dialect classifier and a translator and combine it with LLM outputs from various state-of-the-art proprietary and open-source LLMs. Our experiments showcase significant performance improvements in dialectness-related metrics with RL-tuned LLMs but at the cost of a degradation in MT performance. We hope to investigate this contradictory behavior in future work.

Acknowledgments

This research was supported by Google South & Southeast Asia Research Award 2022. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the funding agency.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

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*, pages 709–728, Bangkok, Thailand. Association for Computational Linguistics.

Abdulkareem Alsudais, Wafa Alotaibi, and Faye Alomary. 2022. Similarities between arabic dialects: Investigating geographical proximity. *Information Processing Management*, 59:102770.

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).

Marta Costa-jussa, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Gonzalez, Prangthip Hansanti, John Hoffman, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630.

Md. Arid Hasan, Prerona Tarannum, Krishno Dey, Imran Razzak, and Usman Naseem. 2024. Do large language models speak all languages equally? a comparative study in low-resource settings. *Preprint*, arXiv:2408.02237.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, Shruti Bhosale, Chenguang Zhu, Karthik Abinav Sankararaman, Eryk Helenowski, Melanie Kambadur, Aditya Tayade, Hao Ma, Han Fang, and Sinong Wang. 2024. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following. *Preprint*, arXiv:2410.15553.

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 *ICLR 2022*.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

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).

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.

Janghwan Lee, Seongmin Park, Sukjin Hong, Minsoo Kim, Du-Seong Chang, and Jungwook Choi. 2024. Improving conversational abilities of quantized large language models via direct preference alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11346–11364, Bangkok, Thailand. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Juan Antonio Perez-Ortiz, Felipe Sánchez-Martínez, Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, Aaron Galiano Jimenez, Antoni Oliver, Claudi Aventín-Boya, Alejandro Pardos, Cristina Valdés, Jusèp Loís Sans Socasau, and Juan Pablo Martínez. 2024. Expanding the FLORES+ multilingual benchmark with translations for Aragonese, aranese, Asturian, and Valencian. In *Proceedings of the Ninth Conference on Machine Translation*, pages 547–555, Miami, Florida, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

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 Romney Robinson, Shahd Abdelmoneim, Kelly Marchisio, and Sebastian Ruder. 2025. ALQASIDA: Analyzing LLM quality and accuracy systematically in dialectal Arabic. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22048–22065, Vienna, Austria. Association for Computational Linguistics.

Shubham Vatsal and Harsh Dubey. 2024. A survey of prompt engineering methods in large language models for different nlp tasks. *Preprint*, arXiv:2407.12994.

Zhiheng Yao, Changsong Yu, and 1 others. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 658–671.