Aladdin-FTI @ AMIYA
Three Wishes for Arabic NLP: Fidelity, Diglossia, and Multidialectal Generation

Jonathan Mutal **Perla Al Almaoui** **Simon Hengchen** **Pierrette Bouillon**
Faculté de traduction et d'interprétation, Université de Genève
iguanodon.ai

Correspondence: first.last@unige.ch

Abstract

Arabic dialects have long been under-represented in Natural Language Processing (NLP) research due to their non-standardization and high variability, which pose challenges for computational modeling. Recent advances in the field, such as Large Language Models (LLMs), offer promising avenues to address this gap by enabling Arabic to be modeled as a pluricentric language rather than a monolithic system. This paper presents Aladdin-FTI, our submission to the AMIYA shared task. The proposed system is designed to both generate and translate dialectal Arabic (DA). Specifically, the model supports text generation in Moroccan, Egyptian, Palestinian, Syrian, and Saudi dialects, as well as bidirectional translation between these dialects, Modern Standard Arabic (MSA), and English. The code and trained model are publicly available.[1]

1 Introduction

The Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2026)[2] introduces the "Arabic Modeling In Your Accent" (AMIYA) shared task (Robinson et al., 2026), a benchmark designed to advance computational modeling of DA. The AMIYA shared task focuses on developing language models that capture the linguistic characteristics of spoken Arabic varieties. Such varieties remain under-represented in existing NLP research and resources (Harrat et al., 2019), although there is a growing interest in studying dialectal varieties and more resources are being created (see e.g. Al-Haff et al. (2022); Momayiz et al. (2024); Al Almaoui et al. (2025)). In this evaluation campaign, systems are assessed on their ability to model DA with respect to dialectal fidelity, understanding, and generation quality using the ALQASIDA benchmark (Robinson et al., 2025).

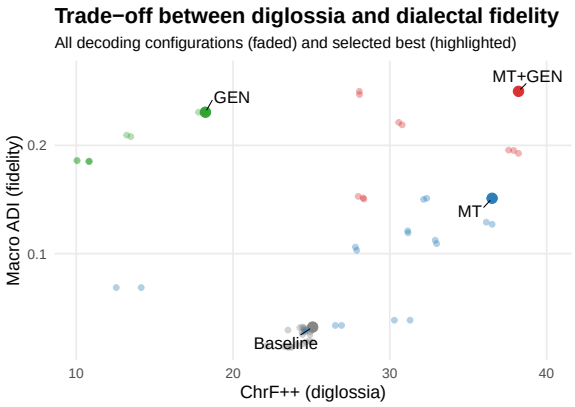

Figure 1: Trade-off between diglossia-sensitive translation accuracy (ChrF++) and dialectal fidelity (Macro ADI2). Each faded point corresponds to a decoding configuration (learning rate × checkpoint), while highlighted points indicate the best configuration selected per model. Instruction-based generation (GEN) favours dialectal fidelity at the expense of diglossia, whereas MT exhibits the opposite behaviour. The combined MT+GEN objective achieves the best overall, improving both fidelity and diglossia.

This paper describes the participation of the team **Aladdin-FTI**, focusing on the closed data track, where the models are only fine-tuned on the official training data provided by the shared task organizers, without the use of additional external corpora. Our approach is based on translation and generation by combining two training objectives: (i) a translation objective aimed at reinforcing diglossic distinctions between MSA and DA, while also preserving semantic adequacy; and (ii) an instruction next-token generation objective designed to produce fluent and linguistically dialectal continuations from partial prompts. By jointly training with these objectives, we seek to have a balance between semantic adequacy and dialectal expressiveness (Robinson et al., 2025). We investigate the complementary roles of

[1] Code: https://github.com/drvenabili/mtfinetune_amiya, models: https://hf.co/collections/unige-fti/aladdin-fti-amiya.

[2] https://sites.google.com/view/vardial-2026

Proceedings of the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 339–351

translation and generation in dialectal Arabic modeling along the following questions:

Q1 What is the impact of fine-tuning for translation on diglossia and dialectal fidelity across Arabic dialects?

Q2 What is the impact of instruction fine-tuning for next word generation on diglossia and dialectal fidelity across Arabic dialects?

Q3 What is the impact of both machine translation (MT) and instruction fine-tuning for next word generation on diglossia and dialectal fidelity across Arabic dialects?

We fine-tune a single large language model under different training settings and evaluate their impact on both diglossia and dialectal fidelity. Our results highlight their distinct yet complementary roles in DA generation.

Our contributions are the following:

- We propose a joint training objective that combines machine translation and instruction-conditioned next-token generation for dialectal Arabic.
- This training enables smaller models to match or even outperform substantially larger baselines in modeling Arabic dialectal variation.

The remainder of this paper is organized as follows: first, Section 2 reviews related work; next, Section 3 presents the methodology; Section 4 describes the experimental setup; Section 5 reports the results; and finally, Section 6 concludes by discussing the limitations of the study.

2 Related Work

Arabic has long been treated as a single homogeneous language, with the majority of resources, benchmarks, and models focusing almost exclusively on MSA. However, this perspective overlooks the deeply diglossic nature of Arabic-speaking communities, in which MSA is rarely a native language and is primarily used in formal and written contexts, while everyday communications take place in regionally and socially diverse dialects (Keleg et al., 2025). These dialects differ substantially from MSA in phonology, morphology, syntax, and lexicon, and each reflects the historical, cultural, and social identities of its speakers (Yushmanov, 1961).

This MSA-centric focus poses significant challenges for NLP systems, as models trained predominantly on MSA tend to normalize or suppress dialectal features when generating or translating dialectal text (Robinson et al., 2023). This gap motivates research into methods that explicitly preserve dialectal characteristics in generated text while maintaining semantic adequacy.

Machine Translation for Dialectal Arabic Prior work has shown that MT provides a natural framework for modeling distinctions between MSA and DA, as translation objectives explicitly condition generation on a source sentence rather than on the target alone (Habash et al., 2013; Zbib et al., 2012). More recent efforts include the fine-tuning of the Kuwain-1.5B small language model for translation from 15 Arabic dialects into Modern Standard Arabic, which achieved high human-rated fluency scores in evaluation studies (Hamed et al., 2025), indicating improved generation quality for dialectal inputs.

Despite these advances, multiple studies report that dialectal MT systems often favour normalised outputs, exhibiting MSA lexical or morphosyntactic choices even when dialectal targets are explicitly specified (Habash et al., 2013; Bouamor et al., 2018; Robinson et al., 2025). While translation-based approaches tend to preserve meaning effectively, they may under-represent dialect-specific variation and linguistic naturalness, a pattern that has also been observed in recent shared-task evaluations (Atwany et al., 2024; Robinson et al., 2023).

Motivated by these findings, we evaluate the impact of machine translation training on diglossia and dialect fidelity, following **Q1**.

Dialectal Text Generation with LLMs Instruction fine-tuning has demonstrated strong performance in controllable text generation (Liang et al., 2024). In dialectal settings, explicit instruction conditioning and the use of dialect tokens have been shown to improve alignment between generated outputs and target dialectal varieties (Barmandah, 2025). However, prior work indicates that such approaches may also introduce generation artifacts when control constraints are emphasized, with potential negative effects on semantic fidelity (Zhang et al., 2023). We therefore evaluate the impact of instruction fine-tuning using next-word prediction on diglossia and dialect fidelity in the context of Arabic dialects (**Q2**).

Combining the Two Tasks Despite advances in both tasks, combining translation-based and generation-based techniques for DA remains underexplored. Prior research on Arabic NLP has begun to explore multi-task learning paradigms, for example, joint modeling of dialect identification and translation to improve MT quality (Khered et al., 2025). There have also been efforts in other settings (e.g. unsupervised MT) to couple translation objectives with language modeling to better preserve fluency (Artetxe et al., 2018). Yet, to our knowledge, no prior work has explicitly jointly optimized a large language model for translation and dialect generation in the Arabic diglossia context. This gap motivates our approach to combine MT and next-word completion (**Q3**).

3 Methodology

In this section, we go through the evaluation protocol (Subsection 3.1) and the training objectives (Subsection 3.2).

3.1 Evaluation Protocol

We followed the evaluation framework proposed by Robinson et al. (2025) to assess two complementary dimensions of DA generation: fidelity and diglossia.

Fidelity This dimension is evaluated using both monolingual and cross-lingual prompts, in which the model is instructed to generate text in a specific DA variety. In such settings, no single gold reference exists, as multiple valid outputs may correspond to the same prompt. Accordingly, fidelity is measured using the Macro ADI2 dialect fidelity score, which assesses whether the generated output is both dialectal and identifiable as the target variety.

Diglossia Diglossia evaluates the model's ability to translate between MSA and DA, reflecting its capacity to differentiate between dialects and MSA. This dimension is assessed through bidirectional translation tasks (MSA→DA and DA→MSA). For these tasks, reference translations are available and performance is measured using ChrF++ (Popović, 2015).

Together, these two dimensions provide a complementary evaluation of dialectal Arabic generation: fidelity focuses on adherence to the target dialect in open-ended settings, while diglossia assesses controlled meaning under reference-based translations.

For each evaluation dimension, we compute the mean score over the different datasets. Specifically, machine translation (corpus-level) and fidelity (sentence-level) are each evaluated on their own dedicated datasets.

3.2 Training Objectives

The two training objectives previously mentioned in section 1 differ not in their optimization procedure but in the constraints imposed by the task formulation. Translation provides a reference to enforce meaning preservation with respect to a source sentence, resulting in a constrained output space. In contrast, dialectal generation is open-ended: for a given instruction and prefix, multiple continuations may be equally valid as long as they are the target dialect (or have linguistic similarities).

Formally, let $\mathcal{D}_{\text{MT}}$ and $\mathcal{D}_{\text{GEN}}$ denote the instruction-formatted datasets used for translation and dialectal generation, respectively. All training examples are represented using an instruction-based format and optimized with a causal language model objective, where only assistant tokens contribute to the loss. The final training objective minimizes a weighted combination of losses over the two datasets:

$$\mathcal{L}_{\text{joint}}(\theta) = \lambda \, \mathrm{E}_{\mathbf{z}\sim\mathcal{D}_{\text{MT}}}[\mathcal{L}(\mathbf{z})] + (1-\lambda) \, \mathrm{E}_{\mathbf{z}\sim\mathcal{D}_{\text{GEN}}}[\mathcal{L}(\mathbf{z})] \quad (1)$$

where $\lambda \in [0, 1]$ controls the relative contribution of translation and generation supervision.

We evaluate the model with $\lambda \in \{0, 0.5, 1\}$. When $\lambda = 0$, the task corresponds to pure generation, whereas when $\lambda = 1$, it corresponds to a MT task.

3.2.1 Machine Translation

For the translation objective, each training example consists of an instruction in English specifying the translation direction (e.g. MSA→DA, DA→MSA, or DA↔English), followed by an assistant response containing the target sentence. After applying the chat template, the model is trained to maximize the conditional likelihood of the assistant tokens given the full preceding context. The template for this task is shown in Table 2.

Prompt examples for translation are illustrated in Table 4.

3.2.2 Instruction Next-Token Generation

For dialectal generation, training examples are formulated as instruction-conditioned sentence completion tasks. Each example provides an explicit

Instruction type	Template	Completion
English instruction	*Complete the sentence starting with these 3 words in <TARGET DIALECT>: <PREFIX>*	*This is the full sentence in <TARGET DIALECT>: <TARGET TEXT IN DIALECT>*
Dialectal instruction	كمّل الجملة وابدأ بالتلات كلمات دول باللهجة <TARGET DIALECT>: <PREFIX>	دي الجملة كاملة باللهجة <TARGET DIALECT>:

Table 1: Templates used for dialectal sentence completion in MADAR training data. Both templates require the model to generate a complete sentence in the target dialect starting from a fixed prefix; only the language of the instruction differs.

Instruction + Source	Target (Reference or Output)	Language Pair
Translate from <SRC> into <TGT>. <SOURCE SENTENCE> Translation:	<TARGET SENTENCE>	<SRC> → <TGT>

Table 2: Template for instructing the translation across Arabic varieties and English.

instruction specifying the target dialect, followed by the first three words of a sentence, which serve as a fixed prefix. The model is trained to generate a complete sentence in the target dialect, starting from this prefix.

Two instruction variants are used. In the first variant, the instruction is formulated in English and specifies both the completion task and the target dialect. In the second variant, the instruction is formulated directly in the target dialect. In both cases, the assistant response contains a full sentence that repeats the provided prefix and continues it in a linguistically coherent manner. Although reference continuations are provided during training, they are not assumed to be unique, as the task may admit multiple valid dialectal realizations for the same prefix. Thus, the training signal encourages the model to learn distributional properties of the target dialect rather than to reproduce a single fixed continuation. The generation templates are provided in Table 1 and examples are shown in Table 5.

4 Experimental Set-Up

4.1 Models

After a hyper-parameter search, we selected SmolLM3-3B to carry out our experiments (Bakouch et al., 2025)[3] as it offers a good balance between model size, performance across tasks, and has been trained with Arabic data.

We instruction fine-tuned SmolLM3-3B to support multiple training objectives and evaluation regimes, selecting the best model according to both MT and next-token generation performance. Periodic evaluation was performed every 1,000 steps using the validation loss and a character-based metric (ChrF++). This design enables a direct comparison of models trained on (a) MT only, (b) MT combined with next-token generation, and (c) next-token generation only, while keeping the evaluation procedure consistent across experimental conditions. We let the reader refer to the model training in Appendix A.1.

To address the research questions, we compare these training configurations against a baseline. The baseline corresponds to SmolLM3-3B with a hyperparameter search over decoding settings ($top\text{-}p \in \{0.1, 0.3, 0.6, 0.9, 1\}$ and $temperature \in \{0.1, 0.3, 0.6, 0.9, 1\}$). We shorten the original template by one third to train our instruction models. We found that both templates provided similar results in preliminary evaluations.

To assess the effect of scaling to a larger model, we replicate the same experimental setting using Llama-3.1-8B-Instruct[4] as the base model, fine-tuned with LoRA (see Appendix A.1). SmolLM-3B builds upon the Llama architecture, with modifications optimized for efficiency and long-context performance, which makes it a suitable point of comparison with Llama-3.1-8B-Instruct. We additionally compare our approach with a larger model to assess its effectiveness relative to models of different sizes, using the best configuration identified during hyperparameter search (refer to Appendix A.4).

4.2 Evaluation Data

We adopted the same evaluation data and protocol described in Al Qasida (Robinson et al., 2025). Our evaluation set comprises both monolingual and crosslingual prompts across multiple DA varieties, designed to support text generation and MT tasks. The crosslingual prompts were drawn from three different collections of LLMs user inputs: (i) a subset of Okapi prompts (Lai et al., 2023) used

[3] https://huggingface.co/HuggingFaceTB/SmolLM3-3B

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

with the Alpaca LLM (Taori et al., 2023), (ii) a collection of ChatGPT prompts obtained via the ShareGPT API, and (iii) a set of human-curated prompts by Marchisio et al. (2024). In addition, our evaluation corpus incorporates both monolingual and bitext sentences from the corpus-6-test-corpus-26-dev split of the MADAR26 (Bouamor et al., 2018), a large-scale parallel resource covering dialects from seven Arab countries and consisting of BTEC-style everyday utterances originally introduced by Takezawa et al. (2007). We further included data from the FLORES200 dev, a multilingual benchmark based on manually translated Wikipedia text, selecting dialectal Arabic subsets representing five major regional varieties (NLLB Team, 2022). Finally, we integrated dialectal Arabic song lyrics from the HABIBI corpus, which spans eight Arab country dialects (El-Haj, 2020).

4.3 Training Data

Following the closed-data track,[5] we used two sources of training data: bilingual data for task (i and ii, refer to Section 3.2.1 and 3.2.2 respectively) and monolingual data for task (ii).

Bilingual Data Our bilingual training data were used to support translation tasks between English, MSA, and DA. For Saudi Arabic, the SauDial corpus (Alanazi et al., 2025) provided parallel data for EN↔DA and DA↔MSA translation. Palestinian Arabic–English parallel data were sourced from the Casablanca corpus (Talafha et al., 2024). For Jordanian Arabic, the JODA corpus (Abandah et al., 2025) was used, offering parallel data between dialectal text and its MSA-corrected version. Syrian Arabic bilingual resources included the UFAL parallel corpus (Krubiński et al., 2023), covering MSA↔DA and DA↔EN translation directions. Moroccan Arabic bilingual data combined several sources: the DODA corpus (Outchakoucht and Es-Samaali, 2024) for EN↔DA translation, and the Atlas training sets (Bounhar and Majjodi, 2025). These data were used to create the machine translation training data (see Section 3.2.1).

Monolingual Data The monolingual training data were compiled from multiple resources covering a wide range of DA varieties. Saudi Arabic monolingual data were obtained from the SDC corpus (Tarmom et al., 2014) as well as from the Saudi Tweets Corpus (Alruily, 2018). The Shami corpus (Abu Kwaik et al., 2018) was used to provide monolingual data for Palestinian, Syrian, and Jordanian varieties, while the MASC corpus (Al-Fetyani et al., 2021) contributed data for Egyptian and Jordanian Arabic. Moroccan Arabic monolingual data were sourced from the Goud training set (Aftiss et al., 2025). Egyptian Arabic monolingual data were enriched using the EDGAD corpus (ElSayed and Farouk, 2020; Hussein et al., 2019), the EDC corpus (Tarmom et al., 2014), and the ASR-EgAr corpus (asr, 2023).

We also incorporated multidialectal monolingual data from the MADAR training set, which includes additional Arabic dialects. These data were used to create the instruction next-token generation data (see Section 3.2.2).

5 Results

Effect of the Task Figure 2 shows the scores with the different training objectives for the diglossia and fidelity tasks. Training with a translation objective improves diglossic scores, as reflected by higher ChrF++ across configurations (**Q1**). While MT also increases dialectal fidelity compared to the baseline, the gains remain limited and highly variable, suggesting that the LLM does not consistently generate the target dialect according to the Macro ADI2 score.

Instruction next-token generation improves dialectal fidelity, yielding the highest Macro ADI2 scores with low variance across configurations. However, this comes at the cost of diglossia, as generation-only models perform poorly on translation tasks, indicating a semantic drift (answering our **Q2**).

Jointly optimizing translation and instruction-based generation objectives yields a balance between diglossic scores and dialectal fidelity. Compared to MT training, the joint model substantially improves Macro ADI2 without degrading ChrF++ (avoiding the semantic shift observed in generation-only models). This indicates that translation and generation objectives provide complementary supervision signals for modeling Arabic dialects.

reprint

Trade-off Between the Tasks Figure 1 illustrates the trade-off between diglossia (measured by ChrF++) and dialectal fidelity, measured by Macro ADI2, across all decoding configurations. Faded points represent individual decoding configurations,

[5] https://sites.google.com/view/vardial-2026/shared-tasks

Model	Diglossia ↑	Fidelity ↑
SmolLM3-3B	22.23	0.003
+ Machine Translation + Instruction (**Constrastive**)	33.65	0.067
Llama-3.1-8B-Instruct	32.99	0.065
+ Machine Translation + Instruction (**Primary**)	35.09	0.233
Command R Arabic	46.60	0.053
GPT-OSS-120B	47.82	0.237

Table 3: Comparison of baseline LLMs in terms of diglossia (ChrF++) and dialectal fidelity (Macro ADI2) across all language. For each model, scores correspond to the best decoding configuration selected across temperature and top-p sampling settings.

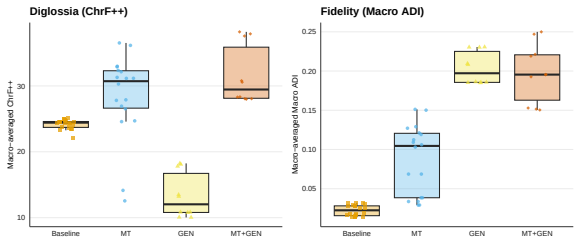

Figure 2: Performance for diglossia (ChrF++) and fidelity (Macro ADI2) across training paradigms. Each boxplot corresponds to a training paradigm (Baseline, MT, GEN, MT+GEN) using SmolLM3-3B, and each point represents a distinct decoding configuration (top_p, temperature, learning rate), with scores macro-averaged over language varieties and test sets.

highlighting the variability induced by learning rate and checkpoint selection (also decoding hyperparameters for the baseline), while the highlighted points correspond to the best-performing configuration selected per model.

The baseline model clusters in the lower-left region of the plot, exhibiting both limited diglossia and weak dialectal fidelity. Models trained exclusively with a MT objective achieve higher ChrF++ scores, indicating stronger diglossia, but remain constrained in dialectal fidelity. In contrast, instruction-based next-token generation (GEN) prioritizes dialectal generation, yielding higher Macro ADI2 scores at the cost of reduced diglossia scores.

The combined MT+GEN model is in the upper-right region of the plot, demonstrating that jointly optimizing translation and instruction-based generation objectives leads to a more favorable balance between semantic adequacy and dialectal expressiveness (**Q3**).

Comparison with Other LLMs To better understand the impact of joint training, we also carried out the same experiments using Llama-3.1-8B-Instruct[6] as the base model, fine-tuned with LoRA.

In addition, we included a 120B-parameter model[7] as a reference point to assess how performance scales with substantially larger model parameters (without any optimization method). We also considered Command R Arabic, a model specifically optimized for translation across multiple Arabic varieties[8], in order to compare our approach against a system designed for Arabic multilingual translation. Taken together, these baselines provide an approximate upper bound on the performance and help contextualize the results of smaller jointly trained models.

Table 3 reports the performance of jointly trained models in comparison with their base model. The automatic evaluation scores indicate that, joint training across tasks leads to consistent improvements in metrics reflecting both diglossia and fidelity scores even for larger model (Llama-3.1-8B-Instruct) with LoRA fine-tuning. These results provide further evidence in support of the benefits of jointly training the model with both task (**Q3**).

Compared to the other models, the much larger GPT-OSS-120B attains the highest scores overall. Despite being substantially smaller, Llama-3.1-8B-Instruct reaches the fidelity score (Macro ADI2) of 0.23, achieving performance comparable to the 120B model. This suggests that dialectal control can be enhanced through supervision strategies rather than model scaling alone.

6 Conclusion

This work presented the submission of team **Aladdin-FTI** to the AMIYA shared task, aiming to model Arabic dialects through a uni-

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

[7] https://huggingface.co/openai/gpt-oss-120b

[8] https://huggingface.co/CohereLabs/c4ai-command-r7b-arabic-02-2025

fied framework that combines translation and instruction-based generation.

Our results show that these objectives provide complementary supervision: translation improves diglossic awareness and semantic adequacy (**Q1**), while instruction-conditioned generation enhances dialectal fidelity (**Q2**). By combining both objectives, we obtain a more balanced model that consistently outperforms single-objective approaches across evaluation dimensions (**Q3**).

Notably, this balance is achieved with a smaller model that competes with larger systems, underscoring the importance of training objective design in dialectal Arabic modeling. These findings support treating Arabic as a pluricentric language.

Future work will focus on refining the balance between objectives, expanding coverage to additional varieties, and conducting human and linguistic evaluations to better assess dialectal naturalness.

Limitations

This study is limited to a single model architecture. While the results are encouraging, further experiments on different model families and scales are needed to assess the generality of the proposed approach. While few-shot and in-context learning approaches may be effective (see e.g. Gao et al. (2021); or Mutal et al. (2025) for use in low-resource settings), they were not considered in this work, as our objective was to keep input prompts compact and limit the number of tokens provided to the model.

Acknowledgments

The computations were performed at the University of Geneva using the Baobab HPC service.

References

2023. ASR-EgArbCSC: An Egyptian Arabic conversational speech corpus. Open-source dataset consisting of 5.5 hours of transcribed Egyptian Arabic conversational speech across nine two-speaker conversations.

Gheith Abandah, Ashraf Suyyagh, Iyad Jafar, Mohammad Abdel-Majeed, Rabie Otoum, Shorouq AlAwawdeh, and Moath Khaleel. 2025. JODA: A dataset of jordanian dialect and erroneous modern arabic sentences coupled with proper MSA and full diacritics.

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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Azzedine Aftiss, Salima Lamsiyah, Christoph Schommer, and Said Ouatik El Alaoui. 2025. Empirical evaluation of pre-trained language models for summarizing moroccan darija news articles. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 77–85.

Perla Al Almaoui, Pierrette Bouillon, and Simon Hengchen. 2025. Arabizi vs LLMs: Can the genie understand the language of aladdin? In *Proceedings of Machine Translation Summit XX: Volume 2*, pages 28–41, Geneva, Switzerland. European Association for Machine Translation.

Mohammad Al-Fetyani, Muhammad Al-Barham, Gheith Abandah, Adham Alsharkawi, and Maha Dawas. 2021. Masc: Massive arabic speech corpus.

Karim Al-Haff, Mustafa Jarrar, Tymaa Hammouda, and Fadi Zaraket. 2022. Curras + baladi: Towards a Levantine corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 769–778, Marseille, France. European Language Resources Association.

Naif Alanazi, Mohammed Al-Batineh, and Hussein Abu-Rayyash. 2025. Saudial: The saudi arabic dialects game localization dataset. *Data in Brief*, 62:111906.

Meshrif Alruily. 2018. Saudi tweets dataset.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium. Association for Computational Linguistics.

Hanin Atwany, Nour Rabih, Ibrahim Mohammed, Abdul Waheed, and Bhiksha Raj. 2024. OSACT 2024 task 2: Arabic dialect to MSA translation. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 98–103, Torino, Italia. ELRA and ICCL.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, and 4 others. 2025. SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3.

Hassan Barmandah. 2025. Saudi-dialect-allam: Lora fine-tuning for dialectal arabic generation. *Preprint*, arXiv:2508.13525.

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).

Abdelaziz Bounhar and Abdeljalil El Majjodi. 2025. Atlaset dataset for moroccan darija: From data collection, analysis, to model trainings. *Hugging Face Blog*.

Mahmoud El-Haj. 2020. Habibi: A multi-dialect arabic song lyrics corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Shereen ElSayed and Mona Farouk. 2020. Gender identification for egyptian arabic dialect in twitter using deep learning models. *Egyptian Informatics Journal*, 21.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. 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 3816–3830, Online. Association for Computational Linguistics.

Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander, and Nadi Tomeh. 2013. Morphological analysis and disambiguation for dialectal Arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 426–432, Atlanta, Georgia. Association for Computational Linguistics.

Mohamed Motasim Hamed, Muhammad Hreden, Khalil Hennara, Zeina Aldallal, Sara Chrouf, and Safwan AlModhayan. 2025. Lahjawi: Arabic cross-dialect translator. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 12–24, Abu Dhabi, UAE. Association for Computational Linguistics.

Salima Harrat, Karima Meftouh, and Kamel Smaili. 2019. Machine translation for Arabic dialects (survey). *Information Processing & Management*, 56(2):262–273.

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. *Preprint*, arXiv:2106.09685.

Shereen Hussein, Mona Farouk, and ElSayed Hemayed. 2019. Gender identification of egyptian dialect in twitter. *Egyptian Informatics Journal*, 20(2):109–116.

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2025. Revisiting common assumptions about Arabic dialects in NLP. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3327, Vienna, Austria. Association for Computational Linguistics.

Abdullah Khered, Youcef Benkhedda, and Riza Batista-Navarro. 2025. A multi-task learning approach to dialectal Arabic identification and translation to Modern Standard Arabic. In *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages*, pages 21–31, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Mateusz Krubiński, Hashem Sellat, Shadi Saleh, Adam Pospíšil, Petr Zemánek, and Pavel Pecina. 2023. Multi-parallel corpus of North Levantine Arabic. In *Proceedings of ArabicNLP 2023*, pages 411–417, Singapore (Hybrid). Association for Computational Linguistics.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and Zhiyu Li. 2024. Controllable text generation for large language models: A survey. *Preprint*, arXiv:2408.12599.

Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. 2024. Understanding and mitigating language confusion in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6653–6677, Miami, Florida, USA. Association for Computational Linguistics.

Imane Momayiz, Aissam Outchakoucht, Omar Choukrani, and Ali Nirheche. 2024. Terjamabench: A culturally specific dataset for evaluating translation models for moroccan darija.

Jonathan Mutal, Raphael Rubino, and Pierrette Bouillon. 2025. Factors affecting translation quality in in-context learning for multilingual medical domain. In *Proceedings of the Tenth Conference on Machine Translation*, pages 161–179, Suzhou, China. Association for Computational Linguistics.

NLLB Team. 2022. No language left behind: Scaling human-centered machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Aissam Outchakoucht and Hamza Es-Samaali. 2024. The evolution of darija open dataset: Introducing version 2. *Preprint*, arXiv:2405.13016.

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.

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 Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. 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.

Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. 2007. Multilingual spoken language corpus development for communication research. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 3, September 2007: Special Issue on Invited Papers from ISCSLP 2006*, pages 303–324.

Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Mohamedou Cheikh Tourad, Rahaf Alhamouri, Rwaa Assi, and 1 others. 2024. Casablanca: Data and models for multidialectal arabic speech recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21745–21758.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Taghreed Tarmom, William Teahan, Eric Atwell, and Mohammad Alsalka. 2014. Compression vs traditional machine learning classifiers to detect code-switching in varieties and dialects: Arabic as a case study. *Journal of Natural Language Processing*.

Nikolai Vladimirovich Yushmanov. 1961. The structure of the arabic language.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59, Montréal, Canada. Association for Computational Linguistics.

Zhihan Zhang, Wenhao Yu, Mengxia Yu, Zhichun Guo, and Meng Jiang. 2023. A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 943–956, Dubrovnik, Croatia. Association for Computational Linguistics.

A LLM Settings and Results

A.1 Instruct Fine-Tuning

We fine-tuned two instruction-tuned models: SmolLM3-3B and Llama-3.1-8B-Instruct. Both models shared the same training configuration. Evaluation and checkpointing were performed every 1,000 steps, and the best-performing model was retained according to ChrF++ and perplexity. ChrF++ evaluation relied on deterministic text generation (temperature 0.0, top-p 1.0) with up to 512 generated tokens, and translation quality was assessed using ChrF++ with a character n-gram size of 64. The best checkpoint was selected by maximizing the macro-averaged ChrF++ score over the full development set.

All models were trained for four epochs with a per-device batch size of 16, gradient accumulation over eight steps (effective batch size 128), and a per-device evaluation batch size of eight. Optimization relied on AdamW with a cosine learning-rate scheduler, a warm-up ratio of 3%, weight decay of 0.01, and gradient clipping with a maximum norm of 1.0. Learning rate values were swept over $\{2 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}, 6 \times 10^{-5}\}$. For reproducibility, we fixed the random seed to 42.

Due to GPU resource constraints, we adopted two different instruct fine-tuning strategies:

SmolLM3-3B We fine-tuned the HuggingFaceTB/SmolLM3-3B model using a custom template to ensure alignment with the training data. Training was conducted in bfloat16 precision with TF32 enabled and gradient checkpointing. No parameter-efficient fine-tuning or quantization was applied for this model.

Llama-8B-Instruct We fine-tuned meta-llama/Meta-Llama-3.1-8B-Instruct using parameter-efficient adaptation with LoRA (Hu

et al., 2021). LoRA was applied with rank $r = 16$, scaling factor $\alpha = 32$, and dropout 0.05, targeting the attention projection layers (q, k, v, o) and the feed-forward layers. No quantization was applied.

A.2 Machine Translation Template

Instruction + Source (English)	Reference Translation (Egyptian Arabic)
Translate from English into Egyptian Arabic. Output only the translation. I wonder if a table is available near the window for seven tonight.	أنا بسأل لو كان فيه ترابيزة جنب الشباك تكون فاضية على الساعة سبعة بليل.
Translate from English into Egyptian Arabic. Output only the translation. I feel chilly and my stomach hurts badly.	حاسس إني بردان ومعدتي واجعاني جامد.
Translate from English into Egyptian Arabic. Output only the translation. Can I invite you out for dinner some time?	ممكن أعزمك على العشا في وقت ما؟
Translate from English into Egyptian Arabic. Output only the translation. The very thing I was looking for.	الحاجة اللي كنت بدوّر عليها.
Translate from English into Egyptian Arabic. Output only the translation. Flight number two zero eight, to Tokyo.	رحلة رقم ميتين وتمانية، لطوكيو.
Translate from English into Egyptian Arabic. Output only the translation. I'd like a typewriter ribbon.	عايز شريط آلة كاتبة.
Translate from English into Egyptian Arabic. Output only the translation. What's the minimum age to do this?	إيه أقل سن ممكن يعمل ده؟
Translate from English into Egyptian Arabic. Output only the translation. Could you call me if you find my bag?	ممكن تكلّمني لو لقيت شنطتي؟

Table 4: Examples of instruction-conditioned machine translation from English to Egyptian Arabic (MADAR-26).

A.3 Instruction Next-Token Generation

Instruction language	Instruction + Prefix	Reference continuation (assistant)
English	*Complete the sentence starting with these 3 words in Egyptian Arabic:* ده قدامك هناك،	دي الجملة كاملة باللهجة المصرية: ده قدامك هناك، يادوبك قدام مكتب استعلامات السياحة.
Dialectal (Egyptian Arabic)	كمّل الجملة وابدأ بالتلات كلمات دول باللهجة المصرية: ده قدامك هناك،	دي الجملة كاملة باللهجة المصرية: ده قدامك هناك، يادوبك قدام مكتب استعلامات السياحة.
English	*Complete the sentence starting with these 3 words in Egyptian Arabic:* عمري ما سمعت	دي الجملة كاملة باللهجة المصرية: عمري ما سمعت عن العنوان ده هنا.
Dialectal (Egyptian Arabic)	كمّل الجملة وابدأ بالتلات كلمات دول باللهجة المصرية: عمري ما سمعت	دي الجملة كاملة باللهجة المصرية: عمري ما سمعت عن العنوان ده هنا.

Table 5: Examples of instruction-conditioned dialectal generation from MADAR training data. The task consists of completing a sentence in a target dialect from a fixed three word prefix. Instructions are provided either in English or directly in the target dialect, while the generation objective remains identical.

A.4 Hyperparameter Search Results

In this section, we show the performance of each model under different decoding hyperparameter settings.

Hyperparameters		SmolLM3-3B		Llama-3.1-8B-Instruct		Command R Arabic		GPT-OSS-120B	
top-p	T	Diglossia	Fidelity	Diglossia	Fidelity	Diglossia	Fidelity	Diglossia	Fidelity
0.1	0.1	21.96	0.003	32.96	0.044	45.34	0.018	47.14	0.200
0.1	0.3	22.00	0.003	32.96	0.044	45.35	0.019	47.39	0.206
0.1	0.6	22.23	0.003	32.97	0.045	45.97	0.019	47.30	0.205
0.1	0.9	22.00	0.003	32.95	0.043	46.42	0.020	47.17	0.227
0.1	1	21.85	0.003	32.83	0.044	46.19	0.021	47.60	0.214
0.3	0.1	21.96	0.003	32.97	0.044	46.42	0.020	47.50	0.207
0.3	0.3	21.93	0.003	32.97	0.044	46.46	0.020	47.55	0.204
0.3	0.6	21.83	0.003	32.86	0.043	46.47	0.021	47.49	0.224
0.3	0.9	21.40	0.003	32.49	0.044	46.52	0.020	47.37	0.216
0.3	1	20.96	0.003	32.20	0.047	46.60	0.053	47.32	0.237
0.6	0.1	21.81	0.003	32.97	0.043	46.23	0.023	47.48	0.188
0.6	0.3	21.80	0.003	32.91	0.045	46.19	0.022	47.53	0.210
0.6	0.6	21.18	0.003	32.31	0.045	46.18	0.022	47.72	0.216
0.6	0.9	18.96	0.002	30.82	0.046	46.12	0.023	47.13	0.208
0.6	1	16.86	0.002	29.69	0.043	46.05	0.021	46.78	0.203
0.9	0.1	21.85	0.002	32.97	0.045	45.19	0.021	47.45	0.209
0.9	0.3	21.68	0.002	32.67	0.044	45.08	0.020	47.48	0.223
0.9	0.6	18.91	0.003	30.81	0.049	45.13	0.021	47.39	0.215
0.9	0.9	13.76	0.002	27.34	0.055	45.23	0.021	47.19	0.197
0.9	1	11.40	0.002	25.21	0.058	45.33	0.020	46.83	0.211
1	0.1	21.78	0.003	32.99	0.045	44.83	0.020	47.50	0.217
1	0.3	21.27	0.002	32.54	0.045	44.75	0.020	47.82	0.213
1	0.6	17.44	0.002	30.06	0.047	44.88	0.021	47.31	0.202
1	0.9	12.09	0.002	25.70	0.048	44.72	0.020	46.97	0.219
1	1	9.78	0.002	23.66	0.065	44.78	0.020	45.68	0.218

Table 6: Decoding performance for different top-p and temperature (T) settings, evaluated with Diglossia (ChrF++) and Fidelity (Macro ADI2).

The table 6 reports the impact of varying decoding hyperparameters p and temperature) on both diglossia (ChrF++) and dialectal fidelity (Macro ADI2) across several models. Overall, performance remains relatively stable for moderate decoding settings, where diglossia scores vary only slightly within each model. However, increasing temperature leads to more diverse but less controlled generation, which often results in degraded diglossia scores, particularly for smaller models such as SmolLM3-3B, whose ChrF++ drops sharply at higher temperature values ($T = 1, top - p = 1, diglossia = 9.78$ diglossia score). OpenGPT-OSS-120B (the largest model) remain more robust, maintaining high diglossia and fidelity scores across most configurations.