SDNLP at AMIYA 2026: Syrian Arabic Dialect Modeling with LoRA

Hasan Alkhder
Sakarya University, Türkiye
hasan.alkhder2@ogr.sakarya.edu.tr

Mohammad Abboush
TU Clausthal, Germany
mohammad.abboush@tu-clausthal.de

Abstract

Dialectal Arabic continues to represent a persistent challenge for contemporary large language models, which are predominantly trained and optimized for Modern Standard Arabic (MSA) and therefore exhibit limited capability when processing colloquial varieties. In this study, a dedicated system developed for participation in the AMIYA shared task focusing on Syrian Arabic is presented. The proposed solution is based on the integration of parameter-efficient fine-tuning through Low-Rank Adaptation (LoRA) with prompt-guided inference, aiming to enhance dialectal adequacy and linguistic naturalness. Rather than emphasizing strict factual precision, the system is deliberately designed to prioritize fluent and authentic Syrian Arabic generation, in accordance with the evaluation principles adopted by the AL-QASIDA benchmark. This design choice reflects a focus on human-perceived language quality and dialectal fidelity, which are central to effective dialect-aware language modeling.

1 Introduction

Dialectal Arabic constitutes the dominant form of daily communication throughout the Arab world; however, it remains largely underexplored within natural language processing research, which has traditionally focused on MSA. As a result, most pretrained large language models exhibit strong performance in formal registers while demonstrating limited capability in generating fluent and natural dialectal language, a shortcoming that is particularly pronounced for Syrian Arabic due to its highly conversational nature and substantial divergence from standardized written forms. This work is conducted within the context of the AMIYA shared task on Syrian Arabic dialect modeling (Robinson et al., 2026).

To address these challenges, the proposed system is guided by a set of explicit design choices tailored to dialect modeling under shared-task constraints. Specifically, we adopt: (i) parameter-efficient fine-tuning via Low-Rank Adaptation (LoRA) to enable dialect specialization without modifying the full model parameters; (ii) prompt-level constraints that explicitly encourage Syrian Arabic realization while discouraging Modern Standard Arabic structures; and (iii) a qualitative evaluation strategy aligned with the AL-QASIDA benchmark (Robinson et al., 2025), prioritizing dialectal fluency and naturalness over strict literal fidelity. Together, these choices frame the task as one of dialect modeling rather than factual question answering, and directly inform the methodological decisions described in the following sections.

From a linguistic perspective, the challenge of dialectal Arabic modeling is further amplified by the phenomenon of diglossia, where speakers naturally alternate between MSA and colloquial varieties depending on context. This linguistic dynamic often leads pretrained models to default to MSA even when prompted for dialectal output, resulting in responses that sound unnatural to native speakers. Addressing this mismatch requires not only suitable training data but also modeling strategies that explicitly encourage colloquial realization and discourage formal constructions. These considerations motivate the design choices adopted in the present system.

2 Task Framing

The AMIYA shared task is approached in this work as a dialect modeling problem rather than a conventional question answering or fact retrieval task. The primary objective is to generate responses that resemble naturally spoken Syrian Arabic as used in everyday interaction, where fluency and dialectal authenticity are prioritized over strict literal faithfulness to the input prompt.

Importantly, deprioritizing strict semantic cor-

Proceedings of the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 359–364
March 29, 2026. ©2026 Association for Computational Linguistics

rectness does not imply unconstrained or nonsensical generation. The system remains grounded in the semantic intent of the input prompt, which serves as an anchor for generation. Rather than enforcing exact word-by-word or structure-preserving fidelity, the model is allowed to paraphrase, generalize, or abstract the input content when doing so results in more natural colloquial expression. In practice, this means that semantic adequacy is preserved, while rigid literal alignment is relaxed.

Nonsensical or hallucinated outputs are mitigated through a combination of instruction-style fine-tuning and constrained inference. During training, the model is exposed to realistic conversational data grounded in coherent dialogue contexts, which implicitly reinforces semantic plausibility. At inference time, prompts explicitly specify both the target dialect (Syrian Arabic) and the expected response format, anchoring generation to the input intent while allowing stylistic flexibility. This combination prevents unconstrained generation while still enabling dialectally natural reformulation.

This task formulation aligns with the evaluation principles of the AL-QASIDA benchmark, which penalizes outputs that drift toward Modern Standard Arabic or overly formal registers, even when they are semantically precise. Consequently, the methodological focus shifts toward controlling stylistic realization rather than enforcing exact semantic reproduction. This trade-off reflects real-world conversational usage, where communicative plausibility and dialectal naturalness often outweigh strict literal accuracy.

3 Data

3.1 Data Source

The training data used in this work consists of professionally produced dubbed television dialogues provided by the professional dubbing company **NIS**. The dataset is external to the official AMIYA shared task release and is therefore used under the Open Track regulations. No data provided by the task organizers or other publicly released shared-task datasets (e.g., UFAL or related corpora) were used in this work.

The original scripts are primarily in English and were translated into Modern Standard Arabic (MSA) and subsequently localized into Syrian Arabic by professional translators as part of the dubbing production workflow. Crucially, the Syrian Arabic content reflects natural spoken usage rather than literal translation, as it is adapted for oral delivery in audiovisual media. This makes the data particularly suitable for dialect modeling tasks that emphasize conversational naturalness over formal written structure.

This localization process yields conversational, informal utterances that closely resemble everyday spoken Syrian Arabic. As the data is generated by professional dubbing practitioners, it exhibits consistent stylistic patterns and strong dialectal authenticity, distinguishing it from automatically generated or crowdsourced resources. In addition, the dialogue-oriented nature of the data supports parameter-efficient adaptation methods such as Low-Rank Adaptation (LoRA) (Hu et al., 2022), which benefit from stylistically coherent training signals.

3.2 Dataset Size and Structure

The dataset comprises approximately **30,000 dialogue utterances**. Each instance corresponds to a single conversational turn and is represented in a structured tabular format with the following fields:

- **character**: the speaker identifier associated with the dialogue turn,
- **gender**: the speaker's gender (male or female),
- **english**: the original English script,
- **arabic_msa**: the corresponding Modern Standard Arabic version,
- **arabic_syrian**: the target Syrian Arabic utterance.

The average utterance length is approximately **9.6 words** in Syrian Arabic, **8.8 words** in MSA, and **9.9 words** in English, reflecting the dialogue-oriented and concise nature of the data. No additional datasets or auxiliary corpora were incorporated during training.

3.3 Suitability for Dialect Modeling

The dataset is inherently dialogue-driven and predominantly informal, reflecting everyday conversational scenarios rather than scripted narration or formal text. This characteristic aligns closely with the objectives of the AMIYA shared task, which prioritizes dialectal fidelity, fluency, and naturalness over strict semantic equivalence.

Moreover, the presence of parallel English and MSA representations implicitly exposes the model to distinctions between formal and colloquial registers. This property is particularly valuable in the context of Arabic diglossia, where pretrained language models often default to MSA even when prompted for dialectal output. By grounding adaptation in professionally localized dialectal utterances, the model is encouraged to internalize colloquial realization patterns that are difficult to capture using automatically generated resources.

3.4 Legal and Ethical Considerations

The data was used with explicit permission for research purposes under an agreement with the data provider and fully complies with the regulations of the AMIYA Open Track. No restricted, confidential, or off-limits resources were incorporated in the development of the proposed system, and no personal or sensitive user information is contained within the dataset.

4 System Overview

The proposed system follows a modular adaptation-based architecture designed specifically for dialect modeling under shared-task constraints. It is built upon a pretrained large language model, which is extended through the integration of a lightweight Low-Rank Adaptation (LoRA) module and a controlled prompt-guided inference strategy. In this configuration, all original parameters of the base model remain frozen, while a small set of additional trainable parameters is optimized to steer the model toward Syrian Arabic generation.

Dialectal specialization is achieved through two complementary mechanisms. First, LoRA-based parameter-efficient fine-tuning allows the model to internalize stylistic and lexical patterns characteristic of Syrian Arabic without altering its general linguistic competence. Second, prompt-guided inference is employed at generation time to explicitly control output structure and register, encouraging concise, single-sentence responses in colloquial Syrian Arabic. Together, these mechanisms enable effective dialect adaptation while maintaining model stability and reproducibility.

The architecture is intentionally designed to be modular. By decoupling the pretrained backbone from the dialect-specific adaptation layer, the system can be easily extended to other Arabic dialects by replacing or retraining only the LoRA components, without requiring full model retraining. This modularity also facilitates experimentation with alternative prompting strategies or constraint formulations, making the approach well suited for iterative refinement in shared-task settings.

An additional motivation for adopting LoRA in this context is its alignment with the computational and practical constraints of shared tasks such as AMIYA. Parameter-efficient adaptation enables targeted dialect specialization with limited resources, avoiding the need for large-scale retraining while preserving the robustness and generalization capabilities of the underlying pretrained model. As a result, the system emphasizes controllable stylistic adaptation rather than architectural complexity.

This system is submitted as the *primary* submission for the Syrian Arabic track of the AMIYA shared task.

5 Model and Training

The proposed system is built upon the **Meta-Llama-3.1-8B** *base* language model (i.e., not the instruction-tuned variant) (Touvron et al., 2023), accessed via the HuggingFace platform.[1] Parameter-efficient fine-tuning is performed using Low-Rank Adaptation (LoRA) (Hu et al., 2022), while keeping all original model parameters frozen.

LoRA adapters are injected into the attention layers of the transformer architecture, specifically targeting the `q_proj`, `k_proj`, `v_proj`, and `o_proj` modules. The adaptation is configured with a rank of **16**, a scaling factor of **32**, and a dropout rate of **0.05**. This configuration enables effective dialect specialization while maintaining training stability and computational efficiency.

Fine-tuning is conducted in a causal language modeling setting using instruction-style input–output pairs derived from the dialogue data. The model is trained for **1.5 epochs** using the AdamW optimizer with a learning rate of 2×10^{-4}. A per-device batch size of **4** is used in conjunction with gradient accumulation over **8** steps, resulting in an effective batch size of **32**. To reduce memory consumption, the model is loaded in 8-bit precision and trained using mixed-precision (FP16). All training procedures strictly comply with the AMIYA Open Track regulations.

[1] `https://huggingface.co/meta-llama/Meta-Llama-3.1-8B`

Submission	Adequacy	Fluency
SDNLP (Primary)	1.124	2.928

Table 1: Official AMIYA 2026 human evaluation results for the SDNLP primary submission on the Syrian Arabic track.

6 Prompting and Inference

During inference, a prompt-guided generation strategy is employed to steer the model toward producing dialectally appropriate outputs. The evaluation prompts are those officially released as part of the AMIYA shared task and distributed by the task organizers under the AL-QASIDA evaluation framework. These prompts typically include explicit instructions such as “Translation:” or “Answer in Syrian Arabic,” and are used in accordance with their original attribution and evaluation guidelines.

To enforce dialectal and structural consistency, the original prompts are augmented at the prompt level with explicit constraints that require the generation of a single sentence in Syrian Arabic while discouraging explanations, meta-commentary, or multi-sentence outputs. These constraints are implemented through prompt formulation rather than decoding-time penalties, ensuring that generation remains anchored to the semantic intent of the input while allowing stylistic flexibility in colloquial realization.

In addition to prompt-level control, a lightweight post-processing step is applied to remove spurious surface-level artifacts. This step consists of simple rule-based filtering to eliminate duplicated tokens, unintended Latin characters, and residual Modern Standard Arabic morphological endings that occasionally appear in generated outputs. No content-level rewriting or semantic modification is performed during post-processing; the procedure is limited to surface-form cleanup to preserve the model’s original generation behavior.

The restriction to single-sentence outputs is motivated by empirical observation rather than a formal ablation study. In preliminary experimentation, unconstrained prompts frequently resulted in longer, explanatory responses that diluted dialectal consistency. By contrast, explicitly constrained prompts yielded outputs that were shorter, more predictable, and more consistently aligned with colloquial Syrian Arabic. This qualitative observation is reflected in the example outputs presented in Section 3.

7 Evaluation

System predictions are generated for all test prompts officially released as part of the AMIYA shared task and provided by the task organizers under the AL-QASIDA evaluation framework. The prompts are processed in their original order without re-ranking or filtering, and system outputs are exported in a comma-separated values (CSV) format that preserves both the original sequence and the exact number of entries, in accordance with the shared task submission guidelines.

All evaluation is conducted centrally by the AMIYA organizers. The evaluation framework combines automatic metrics designed to assess dialectal characteristics with human judgments that focus on perceived fluency and adequacy. No additional test-time adaptation, prompt modification, or post-hoc tuning is applied beyond the inference constraints described in Section 6.

Table 1 presents the human evaluation scores provided by the task organizers. Adequacy measures how well the generated responses align with the intended meaning of the prompt, while Fluency reflects the perceived naturalness and conversational quality of the output in Syrian Arabic. These human judgments play a central role in the AMIYA shared task, where dialectal naturalness and spoken plausibility are prioritized.

Table 2 reports the official automatic evaluation scores. The ADI2 metric reflects overall dialectal performance, while the direction-specific chrF-based scores measure cross-lingual and cross-variety consistency between Dialectal Arabic (DA), English (ENG), and Modern Standard Arabic (MSA). These metrics are intended to capture stylistic and distributional alignment rather than strict semantic equivalence.

Taken together, the automatic and human evaluation results indicate that the proposed system achieves competitive performance within the Syrian Arabic track. While automatic metrics provide supporting evidence of dialectal alignment, the evaluation framework places particular emphasis on human judgments, which better reflect colloquial fluency and dialectal authenticity. The reported scores are consistent with the system’s design goal of prioritizing natural spoken Syrian Arabic over strict semantic optimization.

Submission	ADI2	DA→ENG	ENG→DA	DA→MSA	MSA→DA
SDNLP (Primary)	0.280	36.211	14.813	6.743	15.609

Table 2: Official AMIYA 2026 automatic evaluation results for the SDNLP primary submission on the Syrian Arabic track. Directional scores are reported using chrF/chrF++ metrics as defined in the AL-QASIDA evaluation framework.

Source (English)	Base Model Output	LoRA-Adapted Output
Where are you going now?	يلا قومي على رجلكي وين بدك تروحي لا تروحي من هون	انتي شو عم تعملي هون وين رايحة؟
I don’t want to talk anymore.	انا ما بدي احكي اكتر من هيك خلّيكي فيقي	بدي روح من هون ما عاد بدي حكي معك
What should we do now?	شو بدنا نعمل؟ كيف ح نعيش هيك؟	شو بدنا نعمل هلأ؟
It’s too late now.	انت ما بتسوى شي لاليف خالة رابعة	ما عاد في داعي نضيع وقت هيك

Table 3: Qualitative comparison of output fluency between the base Llama-3.1-8B model and the LoRA-adapted version.

8 Observations and Example Outputs

To assess the impact of dialect-specific adaptation, we conduct a qualitative comparison between the outputs of the LoRA-adapted model and those produced by the original Meta-Llama-3.1-8B base model under identical prompting conditions. This comparison is designed to isolate the effect of parameter-efficient fine-tuning on stylistic and dialectal generation behavior, independent of task formulation or prompt variation.

The evaluation prompts follow a controlled instruction-based format. Each input explicitly requests generation in Syrian Arabic using formulations such as:

```
Respond in Syrian Arabic:
<sentence>
Rewrite in Syrian Arabic:
<sentence>
```

For clarity and readability, Table 3 reports only the core semantic intent of each prompt, while the full instruction templates are described here and were applied consistently during evaluation.

Table 3 presents four representative examples selected from a broader set of prompts used for qualitative analysis. The examples correspond to common conversational scenarios, including refusal, hesitation, decision-making, and expressions of temporal finality. They are not intended to be exhaustive, but rather to illustrate systematic stylistic differences in dialectal realization between the base and adapted models.

Across these examples, the base model frequently exhibits one or more of the following behaviors: (i) introduction of additional narrative or explanatory content not implied by the prompt, (ii) partial drift toward less context-appropriate or more verbose responses, and (iii) overextended utterances that exceed what is typical in spontaneous spoken Syrian Arabic. For instance, in the prompt *“Where are you going now?”*, the base model produces a longer utterance that introduces additional imperatives and discourse elements beyond the pragmatic intent of the input.

In contrast, the LoRA-adapted model consistently produces shorter, more direct utterances that better reflect colloquial Syrian Arabic norms. For example, in *“What should we do now?”*, the adapted model favors a concise spoken formulation that aligns closely with everyday conversational usage, while the base model expands the response with speculative or emotionally loaded content. Across the remaining examples, similar tendencies toward brevity, pragmatic compression, and reduced overgeneration can be observed.

These qualitative differences illustrate a controlled form of generalization rather than semantic degradation. While certain outputs sacrifice literal completeness, they remain contextually appropriate and fluent within a spoken dialogue setting. Importantly, these patterns are consistent with the system's observed behavior under the AL-QASIDA evaluation framework, which emphasizes dialectal fidelity, fluency, and naturalness over strict factual accuracy. The examples therefore serve to concretely illustrate how the proposed system satisfies the primary evaluation criteria of the AMIYA shared task.

9 Conclusion

This work introduces a Low-Rank Adaptation (LoRA)–based system for modeling the Syrian Arabic dialect, developed and submitted within the context of the AMIYA shared task. By conceptualizing the problem as a dialect modeling task and integrating parameter-efficient fine-tuning with prompt-guided inference, the proposed approach attains a high level of dialectal fluency in accordance with the AL-QASIDA evaluation framework. The presented results underscore the suitability and effectiveness of lightweight adaptation strategies for advancing dialectal Arabic language modeling while maintaining computational efficiency.

Limitations

The proposed system is not designed with the primary objective of maximizing factual accuracy or delivering detailed technical explanations. Consequently, when confronted with prompts involving precise dates, named entities, or complex informational content, the generated responses may exhibit a degree of abstraction or generalization. This behavior is an intentional outcome of the adopted design strategy, which prioritizes dialectal authenticity and conversational naturalness over exhaustive semantic completeness.

Concrete examples of this abstraction can be observed in Table 3, where the LoRA-adapted model produces shorter, pragmatically appropriate responses that condense or paraphrase the input prompt rather than reproducing its full literal content. While this may result in reduced semantic specificity, the outputs remain fluent, contextually appropriate, and aligned with the evaluation objectives of the AMIYA shared task.

Acknowledgments

The authors would like to express their gratitude to Mr. Nabeel Al-Dakkak from NIS for granting access to the professional dubbing dialogue data used in this study.

References

Edward J. Hu and 1 others. 2022. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*.

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, Anjali Kantharuban, Otba Alsboul, Salima Lamsiyah, Kelly Marchisio, and Kenton Murray. 2025. Al-qasida: Evaluating dialectal arabic in large language models. In *Proceedings of the AL-QASIDA Shared Task*. Association for Computational Linguistics.

Hugo Touvron and 1 others. 2023. Llama: Open and efficient foundation language models. In *Proceedings of the International Conference on Machine Learning*.