Unicorn at NADI 2025 Subtask 3: GEMM3N-DR: Audio-Text Diacritic Restoration via Fine-tuning Multimodal Arabic LLM

Mohamed Lotfy Elrefai
Ain Shams University
mohamed.lotfy.elrefai@gmail.com

Abstract

We present **GEMM3N-DR**, a multimodal system for NADI 2025 Subtask 3 (*Spoken Arabic Diacritic Restoration*). GEMM3N-DR fine-tunes the *Gemma 3N* LLM via Low-Rank Adaptation (LoRA) using only the official NADI training data, taking both audio and undiacritized text as input and generating fully diacritized output. We apply data augmentation with the nlpaug and the CATT diacritization model. At inference time, we use a structured Arabic instruction and 7-shot examples. Our system achieved a Word Error Rate (WER) of **64%** and Character Error Rate (CER) of **15%** on the hidden test set, ranking **in 2nd place** in the competition. We provide a detailed analysis of model performance, including common error types such as hallucination and incomplete outputs.

1 Introduction

Arabic diacritic restoration is the task of predicting short vowels and other diacritic marks that are omitted in standard Arabic orthography. The problem becomes more challenging in spoken domains, especially for dialectal Arabic, where morphology and phonetics diverge from Modern Standard Arabic (MSA). This task has strong implications for improving readability, ASR post-processing, TTS, and educational tools.

The **NADI 2025 Subtask 3** (Talafha et al., 2025)focuses on diacritic restoration of spoken Arabic dialects using both audio and text. Our approach, **GEMM3N-DR**, leverages the multimodal *Gemma 3N* LLM, adapting it to this task with Low-Rank Adaptation (LoRA) fine-tuning, multi-example prompting, and audio-text fusion.

Our main contributions:

- First application of *Gemma 3N* to spoken Arabic diacritization.
- LoRA-128 fine-tuning with nlpaug-based audio augmentation.
- Use of CATT (Alasmary et al., 2024) predictions as auxiliary inputs for robust training for unlabeled samples, like augment part.
- Structured 7-shot Arabic prompts for inference, reducing WER from 79.05 to 69.05 on devset.

2 Background

The **NADI 2025 Subtask 3: Diacritic Restoration of Spoken Arabic Dialects** (Talafha et al., 2025) challenges participants to restore full Arabic diacritics given an undiacritized transcript and its corresponding speech signal. The task is motivated by the practical need to improve the usability of automatic speech recognition (ASR) outputs, assist language learners, and enhance downstream applications such as text-to-speech (TTS) synthesis.

Task Setup. Participants are given a set of audio-transcript pairs, where transcripts are stripped of diacritics. The goal is to produce fully diacritized text. An example is shown in Table 1.

Input (Undiacritized)	Target (Diacritized)
هذا كتاب جديد	هٰذَا كِتَابٌ جَدِيدٌ

Table 1: Example of task input/output for NADI Subtask 3.

Text-Based Diacritization. Restoring diacritics for written Modern Standard Arabic (MSA) is a well-established problem. Early approaches relied on hand-crafted morphological rules and analyzers, as seen in systems like Madamira (Pasha et al., 2014) and Camelira (Obeid et al., 2022). The field has since evolved through statistical methods to modern deep learning architectures. These include neural sequence-to-sequence models, bidirectional LSTMs followed by Conditional Random

Proceedings of The Third Arabic Natural Language Processing Conference, pages 767–773
November 8-9, 2025 ©2025 Association for Computational Linguistics

Fields (CRF) (Al-Thubaity et al., 2020), and more recently, specialized character-level transformers like **CATT** (Alasmary et al., 2024). A significant recent contribution is **Sadeed** (Aldallal et al., 2025), a decoder-only language model specifically pre-trained and fine-tuned on diverse Arabic corpora. By focusing on high-quality diacritized data, Sadeed demonstrates that specialized models can perform better than general-purpose architectures like CATT, representing a strong benchmark for text-based diacritization.

Audio-Assisted Diacritization. In contrast, the use of audio information to assist in diacritization is a developing field. Text-based models experience a significant performance drop when applied to speech transcripts due to the shift of the domain to the informal spoken language and the prevalence of dialectal variants (Shatnawi et al., 2023). This inadequacy is well documented, with studies showing that speech models trained on gold diacritized data outperform those using text-restored transcripts, highlighting the need for speech-specific solutions (Aldarmaki and Ghannam, 2023).

Pioneering work by (Vergyri and Kirchhoff, 2004) first explored using acoustic information for this task decades ago. Only very recently has this idea been revisited with modern deep learning. Research has branched into complementary approaches: one line of work, exemplified by (Shatnawi et al., 2023), uses a cascaded framework where a fine-tuned Whisper ASR model generates diacritized transcripts to enhance a text-based restoration model. Another approach moves **Beyond Orthography** to directly recover short vowels and dialectal sounds. (Kheir et al., 2024) proposed a novel framework utilizing discrete codes to represent dialectal variability, showing strong performance with limited data and introducing a new dialectal benchmark dataset.

While these methods show promise, they represent disconnected solutions. The former is a cascaded, two-stage pipeline, and the latter focuses on a specific acoustic modeling approach. Our work unifies these directions by proposing a single, end-to-end **multimodal LLM**. Unlike cascaded systems, our model jointly processes raw audio and text signals to directly disambiguate homographs and dialectal variants, effectively bridging the gap between high-quality text diacritization and the challenges of the speech domain.

2.1 Dataset

We used the dataset from the NADI 2025 Shared Task (Subtask 3: Automatic Speech Diacritization) (Talafha et al., 2025), which provides parallel audio-transcript pairs. We participated in the **closed** track is a competition requiring participants to use only the provided resources for a fair comparison. The dataset encompasses a wide range of Arabic varieties and recording conditions, including Dialectal (DIA), Modern Standard (MSA), Classical (CA), and Code-Switched (CS) Arabic.

The training data is composed of two distinct parts:

- **Diacritized Data:** Transcripts with fully vocalized gold standard diacritics (e.g. بَعْدَ الرُّسُومِ الْمُسِيئَةِ لِلنَّبِيِّ ص عَامَ أَلْفَيْنِ وَ سِتَّة تَحَوَّلَتْ حَيَاةُ الْمُسْلِمِينَ فِي الدَّانِمَارْك فَأَصْبَحُوا يَتَمَتَّعُونَ بِكَثِيرٍ مِن لْحُقُوقِ).
- **Non-Diacritized (Augment) Data:** Raw transcripts without diacritics, containing dialectal and code-switched content (e.g., فحين مثلًا أقدر أقول اللي كان يدفعك إنك إنت ما تستسلم هذا التساؤل اللي بينك وبين نفسك إن أنا Senior director.)

2.2 Dataset Statistics

The training set is composed of over 85K sentences drawn from various constituent datasets, each representing a specific Arabic variety. The composition of these datasets is detailed in Table 6. To ensure consistency and quality, samples containing fewer than three words were removed, and punctuation was eliminated from all texts. The resulting dataset consists of 57K samples for training and 1.5K for development (dev), as summarized in Table 2. The training data is further divided into a fully diacritized portion (train) and a partially diacritized portion used for augmentation (augment).

Split	**#Utterances**	**Hours**	**Avg. Dur. (s)**
Train	51517	88.89	6.21
Augment	6087	14.11	8.34
Dev	1580	1.48	3.36
Test	365	0.79	7.83

Table 2: Overall statistics of the NADI 2025 Subtask 3 dataset splits after filtering.

3 System Overview

Our diacritization system is built upon the **Gemma 3N** instruction-tuned language model, which we adapt for the task of Arabic text diacritization using a combination of data augmentation and parameter-efficient fine-tuning. The complete pipeline, from data preparation to final inference, is illustrated in Figure 1 and detailed in the subsequent subsections.

3.1 Augmentation

To enhance the robustness and generalization of our model, we employed a dual-strategy data augmentation approach to effectively increase the size of our training corpus.

- **Audio Augmentation:** Applied a diverse set of audio transformations (pitch shift, noise addition, cropping, speed alteration) using `nlpaug` to enhance acoustic variability, effectively doubling the training data.

- **Text Diacritization:** Utilized the **CATT** model to generate pseudo-labels for non-diacritized text from augmented audio.

3.2 Fine-Tuning

We adapted the pre-trained **Gemma 3N** model to the diacritization task using **LoRA** (Hu et al., 2022).

- **Base Model**: `gemma-3n-E4B-it`

- **PEFT Method**: LoRA

- **Target Modules**: Applied to the key projection matrices within the transformer architecture, specifically targeting both the standard attention mechanisms and audio-specific layers. The targeted modules include:
  - **Attention Projections**: `q_proj`, `k_proj`, `v_proj`, `o_proj`.
  - **Feed-Forward Projections**: `gate_proj`, `up_proj`, `down_proj`.
  - **Audio-Specific Projections**: `post`, `linear_start`, `linear_end`, `embedding_projection`.

- **Hyperparameters**: **Rank** (r): 128, **Alpha** (α): 16, **Dropout**: 0.0

- **Training Setup**: We used the `SFTTrainer` (Supervised Fine-Tuning Trainer).

- **Checkpoint**: The best-performing model was selected from **checkpoint 16500** for final evaluation.

3.3 Inference

At inference time, the model diacritizes raw, non-diacritized Arabic text and audio using a structured prompt-based approach.

- **Prompting**: A fixed **7-shot examples prompt** is used at inference time, consisting of instructions and example pairs.

- **Decoding Parameters**: **Temperature** = 0.001, **Top-*p*** = 1.0 , **Max New Tokens** = 256.

- **Non-Arabic Word Preservation:** Non-Arabic words remain unmodified, maintaining the original sentence structure and ensuring the integrity of code-switched content.

4 Experimental Setup

Our investigation is divided into three primary phases: (1) establishing a baseline performance without any fine-tuning, (2) evaluating parameter-efficient fine-tuning using LoRA, and (3) exploring the effect of increasing the few-shot examples during inference time. All models were evaluated and reported in word error rate (WER% and CER%), where a lower score indicates better performance.

4.1 Baseline Without Fine-tuning

The initial phase establishes a performance baseline for the pre-trained Gemm3n model under two input conditions: using both text and audio data, and using text data alone.

4.2 Fine-tuning With LoRA Parameters

The second phase explores parameter-efficient fine-tuning using LoRA. We experimented with two distinct configurations: a standard LoRA setup with a rank of 8, trained for 5,000 steps, and a more powerful setup combining a high LoRA rank (128) with the 7 few-shot examples identified in the next phase. This aims to quantify the gains from combining advanced fine-tuning techniques with effective prompting.

4.3 Best Fine-tuning Model With Few-Shot Examples At Inference Time

In the Final phase, we investigated the impact of increasing the few-shot examples during inference time on the model (denoted as Gemm3n_F) with a varying number of few-shot examples. The model was evaluated on the development (dev) set, specifically with 3 and 7 examples, to determine if an increased number of few-shot examples improves generalization. The best-performing model checkpoint (at step 16500) was selected for final evaluation to ensure optimal results.

4.3.1 Training Fine-tuning Prompt

We used the following prompt format for training:
System Prompt:
أنت مدقق لغوي لديك ملف صوتي وترى الكلام
المكتوب لتخرج أفضل تَشْكِيل للكلمات العربية
فقط وأترك الكلام غير العربي كما هو كالإنجليزية
والفرنسية على سبيل المثال
User Prompt:
قم بالمراجعة للنص مع المحافظة على نفس عدد
الكلمات ولا تخرج كلمات جديدة فقط أضف
التشكيلات للكلمات العربية
+ Audio Input
+Text Input without diacritic
Assistant Response:
Label Text without diacritic

4.3.2 Inference Time

We have used similar prompt used in training finetuning with n examples as a few shots.we created a method to determine if the word is non arabic and perserving the position
System Prompt:
أنت مدقق لغوي، لديك ملف صوتي والحروف
المكتوبة، أخرج التشكيل الأمثل لكافة الحروف
العربية
User Prompt:
رجاءً أضف التشكيل لكل حرف من الحروف
العربية في الجملة التالية:

مثال ١:
ذهب محمد إلى المدرسة
ذَهَبَ مُحَمَد إِلَى الْمَدْرَسَةِ .. n examples
النص : Text Input without diacritic
+Audio Input

5 Results and Discussion

The results from our comprehensive experiments are presented below, revealing clear trends regarding the impact of input modalities, few-shot prompting, and parameter-efficient fine-tuning with LoRA.

5.1 Baseline Performance Without Fine-tuning

The initial baseline performance of the pre-trained Gemma model is summarized in Table 3. Contrary to the expectation that multimodal input would enhance performance, the model performed significantly better when processing text-only inputs. The Word Error Rate (WER) for text-only inputs was 71%, a substantial 13 percentage point improvement over the 84% WER achieved with combined text and audio inputs. This result suggests that the pre-trained model may not be effectively leveraging the audio modality; the audio features might be introducing noise or the model's fusion mechanism may be suboptimal for this specific task in a zero-shot setting. It's shown from the table 3 result that the audio representations don't align cleanly with the text task. The model treats irrelevant variations (background noise, accents, prosody) as meaningful, reducing performance. Text-only models are more robust because they avoid this noisy modality.

Model	WER%	CER%	Input Modality
Gemm3n	84	34	Text + Audio
Gemm3n	71	23	Text Only

Table 3: WER and CER on the test set performance with different input modalities without any fine-tuning.

5.2 Impact of LoRA Rank on Performance

Our experiments with LoRA yielded the most significant performance gains, as detailed in Table 4. Applying a standard LoRA configuration (rank=8) for 5,000 steps provided a marginal improvement, reducing the test WER to 82% from the multimodal baseline of 84%. The most effective strategy overall was the combination of a high-capacity LoRA fine-tuning (rank=128) and the 7 few-shot examples are identified in Section 5.3. This configuration achieved a test WER of 64%. This represents a dramatic 20 percentage point improvement over the original multimodal baseline.

Model and LoRA rank	WER%	CER%
Gemm3n_F rank=8	82	35
Gemm3n_F rank=128 + 7-shots	64	15

Table 4: Test set WER and CER after fine-tuning with different LoRA configurations.

5.3 Impact of Few-Shot Examples During Inference

We investigated the impact of providing a varying number of few-shot examples at inference time to the best finetuned model (Gemm3n_F). The results, presented in Table 5, show a clear positive correlation between the number of examples and model performance. Using 7 examples during inference yielded a development set WER of 69.05%, outperforming the configuration with only 3 examples, which achieved a WER of 73.21%. This demonstrates that the model can effectively decrease the hallucination and improve its generalization on the development set.

Model	Few-shots	WER%	CER%
Gemm3n_F	3	73.21	23.22
Gemm3n_F	7	69.05	20.84

Table 5: Deve set WER and CER for the best finetuned model (Gemm3n_F, checkpoint 16500) with a varying number of few-shot examples provided at inference time.

5.4 Analysis of Common Error Types

A qualitative analysis of the predictions from the best-performing model (Gemm3n_F) reveals two primary and distinct error patterns, as illustrated in Table 7 and Table 8.

Table 7 demonstrates the first error type: character-level hallucination and modification. Here, the model does not merely add diacritics but incorrectly alters the base characters themselves (e.g., generating بكَم instead of the reference بكَام). This suggests the model's phoneme-to-grapheme conversion is error-prone, leading to changes in the core lexical items, which is a critical failure mode for a transcription task.

Conversely, Table 8 highlights the second error type: inconsistent diacritization due to data sparsity. For words or syntactic structures likely underrepresented in the training data, the model defaults to a safe, undiacritized output (e.g., أبحث instead of أَبحَثُ). This indicates a failure in generalization and a lack of confidence on unfamiliar patterns.

Our experiments, particularly the improvement from 73.2% to 69.05% WER on the development set by incorporating more diverse few-shot examples, point towards effective strategies to mitigate the observed errors. The performance gain achieved by using examples from different dialects and domains (e.g., formal MSA, Egyptian Arabic, Moroccan Arabic) is significant. This approach directly addresses the error of inconsistent diacritization by providing the model with a richer, more representative context of the task during inference. It acts as a dynamic, in-context learning signal that guides the model towards the desired output style and complexity.

6 Conclusion

In conclusion, our experiments demonstrate that while the pre-trained model struggles with raw multimodal inputs, its performance can be significantly enhanced through a dual approach: (1) parameter-efficient fine-tuning with a high-rank LoRA to adapt the model to the task, and (2) leveraging few-shot examples during inference to provide contextual guidance.

For reproducibility, the implementation and code are available[1] at `Unicorn at NADI 2025 Subtask 3`.

References

Ahmed Amine Ben Abdallah, Ata Kabboudi, Amir Kanoun, and Salah Zaiem. 2023. Leveraging data collection and unsupervised learning for code-switched tunisian arabic automatic speech recognition. *Preprint*, arXiv:2309.11327.

[1] Full repository URL: https://github.com/MohamedElrefai/GEMM3N-DR-NADI-2025-Subtask-3

Maryam Khalifa Al Ali and Hanan Aldarmaki. 2024. Mixat: A data set of bilingual emirati-English speech. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 222–226, Torino, Italia. ELRA and ICCL.

Abdulmohsen Al-Thubaity, Atheer Alkhalifa, Abdulrahman Almuhareb, and Waleed Alsanie. 2020. Arabic diacritization using bidirectional long short-term memory neural networks with conditional random fields. *IEEE Access*, 8:154984–154996.

Faris Alasmary, Orjuwan Zaafarani, and Ahmad Ghannam. 2024. Catt: Character-based arabic tashkeel transformer. *arXiv preprint arXiv:2407.03236*.

Zeina Aldallal, Sara Chrouf, Khalil Hennara, Mohamed Motaism Hamed, Muhammad Hreden, and Safwan AlModhayan. 2025. Sadeed: Advancing arabic diacritization through small language model. *arXiv preprint arXiv:2504.21635*.

Hanan Aldarmaki and Ahmad Ghannam. 2023. Diacritic recognition performance in arabic asr. *arXiv preprint arXiv:2302.14022*.

Khalid Almeman, Mark Lee, and Ali Abdulrahman Almiman. 2013. Multi dialect arabic speech parallel corpora. In *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, pages 1–6.

Injy Hamed, Ngoc Thang Vu, and Slim Abdennadher. 2020. ArzEn: A speech corpus for code-switched Egyptian Arabic-English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4237–4246, Marseille, France. European Language Resources Association.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Yassine El Kheir, Hamdy Mubarak, Ahmed Ali, and Shammur Absar Chowdhury. 2024. Beyond orthography: Automatic recovery of short vowels and dialectal sounds in arabic. *arXiv preprint arXiv:2408.02430*.

Ajinkya Kulkarni, Atharva Kulkarni, Sara Abedalmon'em Mohammad Shatnawi, and Hanan Aldarmaki. 2023. Clartts: An open-source classical arabic text-to-speech corpus. In *2023 INTERSPEECH*, pages 5511–5515.

Ossama Obeid, Go Inoue, and Nizar Habash. 2022. Camelira: An arabic multi-dialect morphological disambiguator. *arXiv preprint arXiv:2211.16807*.

Arfath Pasha, Mohamed Al-Badrashiny, Mona T Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *Lrec*, volume 14, pages 1094–1101.

Sara Shatnawi, Sawsan Alqahtani, and Hanan Aldarmaki. 2023. Automatic restoration of diacritics for speech data sets. *arXiv preprint arXiv:2311.10771*.

Bashar Talafha, Hawau Olamide Toyin, Peter Sullivan, AbdelRahim Elmadany, Abdurrahman Juma, Amirbek Djanibekov, Chiyu Zhang, Hamad Alshehhi, Hanan Aldarmaki, Mustafa Jarar, Nizar Habash, and Muhammad Abdul-Mageed. 2025. Nadi 2025: The first multidialectal arabic speech processing shared task. In *The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou. Association for Computational Linguistics.

Hawau Toyin, Rufael Marew, Humaid Alblooshi, Samar M. Magdy, and Hanan Aldarmaki. 2025. ArVoice: A Multi-Speaker Dataset for Arabic Speech Synthesis. In *Interspeech 2025*, pages 4808–4812.

Dimitra Vergyri and Katrin Kirchhoff. 2004. Automatic diacritization of arabic for acoustic modeling in speech recognition.

A Appendix

1.1 Figures

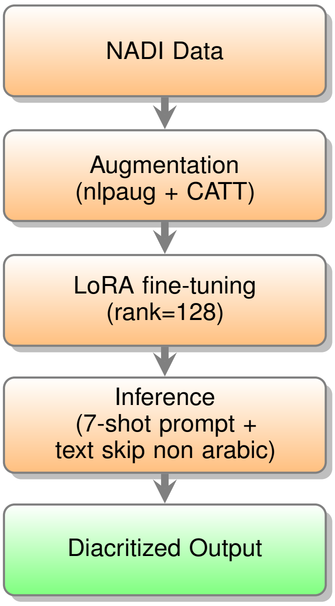
NADI Data
Augmentation (nlpaug + CATT)
LoRA fine-tuning (rank=128)
Inference (7-shot prompt + text skip non arabic)
Diacritized Output

Figure 1: End-to-end pipeline for the diacritization system.

1.2 Tables

Dataset	Type	Diacritized	# Sentences
MDASPC (Almeman et al., 2013)	Multi-dialectal	True	60,677
TunSwitch (Abdallah et al., 2023)	Dialectal, CS	True	5,212
ClArTTS (Kulkarni et al., 2023)	Classical (CA)	True	9,500
ArVoice (Toyin et al., 2025)	MSA	True	2,507
Subtotal		**True**	**77,896**
ArzEn (Hamed et al., 2020)	Dialectal, CS	False	3,344
Mixat (Al Ali and Aldarmaki, 2024)	Dialectal, CS	False	3,721
Subtotal		**False**	**7,065**
Total			**84,961**

Table 6: Breakdown of the constituent datasets within the NADI 2025 original training set.

Reference	Prediction
هُوَّ البُوفِيه المَفتوح بِكَام	هُوَ الْبُوفِيه الْمَفْتُوح بِكَم ؟
هُوَّ فِيه رُسُوم لِلخِدمَه	هُوَ فِيه رُسُومٌ لِلْخِدْمَة

Table 7: Comparison of Model gemm3n_F 7 shots hallucination output compared to Reference by modifying the input text

Reference	Prediction
وَاشُنطُن دِي سِي	واشنطن دي سي
أَنَا أَبحَثُ عَن	أنا أبحث عن

Table 8: Comparison of Model gemm3n_F 7 shots Output against Reference (Undiacritized Samples)