Fine-Tashkeel at KSAA-2026: A Comprehensive Evaluation of Seq2Seq and Multimodal Approaches for Automatic Diacritization of Arabic Speech Dictation

Hassan Barmandah[1,2], **Fatimah Emad Eldin**[1,3], **Omer Nacar**[1,4], **Wareef Alzubaidi**[1,5]
[1] NAMAA Community
[2] Umm Al-Qura University [3] Trouve Labs [4] Tuwaiq Academy [5] King Abdulaziz University
s445001043@uqu.edu.sa, Fatimah@trouve.works,
o.najar@tuwaiq.edu.sa, walzubaidi0014@stu.kau.edu.sa

Abstract

This paper presents the Fine-Tashkeel system for Task 2 of the KSAA-2026 Shared Task on Automatic Diacritization of Speech Dictation. Diacritization of speech-derived Arabic text poses challenges due to dialectal variation, morphological ambiguity, and the absence of acoustic cues in text-only pipelines. Our approach treats diacritization as a character-level sequence-to-sequence task, mapping undiacritized text directly to its fully diacritized form. We evaluate 18 models spanning text-only, ASR-augmented, and fine-tuned configurations, finding that text-only Seq2Seq approaches outperform off-the-shelf multimodal models—a gap we attribute to task mismatch in generic ASR systems rather than an inherent audio limitation. Our best submission, using zero-shot inference without task-specific training, achieved a Diacritic Error Rate (DER) of 10.56%, Word Error Rate (WER) of 34.47%, and Sentence Error Rate (SER) of 79.88%, ranking 5th out of 7 teams. Per-nationality error analysis reveals significant dialectal variation (Egyptian 3.70% vs. Algerian 13.73% DER), and diagnostic analysis confirms that case endings and vowel ambiguity are the primary bottlenecks. Code and evaluation scripts are publicly available.

Keywords: Arabic Diacritization, Sequence-to-Sequence, Speech Dictation, KSAA-2026, Tashkeel, Multimodal

1. Introduction

The KSAA-2026 Shared Task introduces a multimodal benchmark for transforming raw Arabic speech transcripts into fully diacritized text (Al Wazrah et al., 2026). Diacritic restoration remains a core challenge in Arabic NLP: a single undiacritized word (*rasm*) can have multiple valid readings depending on morphosyntactic context. Speech dictation compounds this: ASR systems produce undiacritized text, while standard diacritization models cannot leverage acoustic or dialectal cues.

We propose **Fine-Tashkeel**, a Seq2Seq system (Al-Rfooh et al., 2023) that maps undiacritized character sequences to fully diacritized counterparts. Our key outcomes:

- **Ranking:** 5th out of 7 teams (DER 10.56%, WER 34.47%).
- **Systematic Comparison:** We evaluate 18 models across text-only, ASR+Text, and fine-tuned configurations; text-only Seq2Seq consistently outperforms off-the-shelf multimodal systems due to task mismatch, not an inherent limitation of audio.
- **Error Analysis:** Per-nationality reveals dialect-dependent variation: Egyptian Arabic benefits from representation, while Gulf and Maghrebi show higher rates.

Our code and evaluation scripts are publicly available.[1]

2. Background and Related Work

Arabic diacritization has evolved from rule-based systems to neural architectures. Transformer-based models (Vaswani et al., 2017) enabled sequence classification approaches by Fadel et al. (2019) and Barqawi and Zerrouki (2017). Speech-aware diacritization adds complexity: transcripts lack structural punctuation, contain dialectal variations defying MSA morphological rules, and multilingual models like mT5 (Xue et al., 2021) have shown promise for character-level generation.

Al-Rfooh et al. (2023) proposed a byte-level Seq2Seq approach that generates the full diacritized string end-to-end. For speech-aware diacritization, Shatnawi et al. (2024) demonstrated that leveraging Whisper ASR outputs as auxiliary input to a Transformer diacritization model improves performance over text-only baselines on speech data, establishing the framework adopted as the official baseline for this shared task.

3. Task Setup and Dataset

Systems must process speech audio alongside undiacritized transcripts to generate fully diacritized

[1] https://github.com/HasanBGit/KSAA2026-Fine-Tashkeel

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 234–246
May 11, 2026.

text (Al Wazrah et al., 2026). Hosted on CodaBench, the task prohibits (a) external data beyond the official training set and (b) large language models; only small-scale models are permitted. The target language covers Modern Standard Arabic (MSA) and multiple regional dialects.

3.1. Dataset Statistics

The dataset, collected via the King Salman Global Academy's VoiceWall platform (Al Wazrah et al., 2026), comprises approximately five hours of Arabic speech. Table 1 summarizes the splits and their key characteristics.

Split	Samples	Duration	Nationalities
Train	2,327	∼4.5h	9
Dev	260	∼0.5h	9
Test	328	∼0.7h	9

Table 1: Dataset statistics across splits.

The test set spans nine nationalities: Egypt (173), Saudi Arabia (85), Qatar (34), Sudan (10), Gaza (13), Bahrain (8), Kuwait (2), Syria (1), and Palestine (2). Algeria appears in dev but not test; Gaza in test but not dev—coverage shifts compound dialectal generalization across phonological and morphological systems.

4. System Overview

The pipeline, illustrated in Figure 1, treats diacritization as direct translation from unvoweled to fully voweled text.

4.1. Model Architecture

Our system is built on the `Fine-Tashkeel` model (Al-Rfooh et al., 2023), a Seq2Seq Transformer that, instead of classifying diacritics over a frozen string, frames diacritization as a translation task:

$$Y = \text{Seq2Seq}(X) \tag{1}$$

where $X = \{x_1, x_2, \ldots, x_n\}$ is the sequence of undiacritized characters, and $Y = \{y_1, y_2, \ldots, y_m\}$ is the generated sequence containing both base characters and their corresponding diacritic marks.

This approach handles morphological fusions and can correct transcript normalization errors.

4.2. Inference Pipeline

We employ greedy search optimized for character-level precision (Appendix 10) to prioritize the most probable morphological sequence and prevent hallucinated characters, with a maximum sequence length of 1024 tokens.

5. Experimental Setup

5.1. Models Evaluated

We conducted a systematic evaluation of 18 models across four categories:

1. **Text-only Seq2Seq:** Fine-Tashkeel (Al-Rfooh et al., 2023) (our primary system), Shakkelha (Fadel et al., 2019), Shakkala (Barqawi and Zerrouki, 2017), mT5-base (Xue et al., 2021)

2. **Text-only Classifiers:** ByT5 (glonor, 2024), FLAN-T5 (Chung et al., 2024), Qwen-1.5 (Bisher, 2025), CATT (Alasmary et al., 2024), Mishkal (Zerrouki, 2020), CAMeL-MLE (Obeid et al., 2020)

3. **ASR+Text Multimodal:** Seamless M4T (Seamless Communication et al., 2023), ArTST (Toyin et al., 2023), Whisper variants (Radford et al., 2023; tarteel-ai, 2022; MaddoggProduction, 2026)

4. **Fine-tuned Models:** Tashkeel-700M (Etherll, 2025), ByT5 fine-tuned (glonor, 2024), Whisper-Tashkeel fine-tuned (WajeehAzeemX, 2024)

The full list of all 18 models with their blind test scores is provided in Table 5 (Appendix 2). In addition to inference-based evaluation, we explored several training strategies—including Fine-Tashkeel fine-tuning, curriculum learning, LoRA adaptation, ensemble voting, and post-processing rules—detailed in Appendix 18. None surpassed the zero-shot Fine-Tashkeel baseline, motivating our final submission choice.

5.2. Evaluation Metrics

System performance is evaluated using three complementary metrics:

- **Diacritic Error Rate (DER):** The proportion of incorrectly predicted diacritics at the character level, computed as $(S_c + D_c + I_c)/N_c$, where S_c, D_c, and I_c denote substitutions, deletions, and insertions respectively.

- **Word Error Rate (WER):** The primary metric. A word is marked incorrect if it contains *any* diacritic error, making it highly sensitive to case endings (*i'rab*).

- **Sentence Error Rate (SER):** A strict metric where a sentence is incorrect if any diacritic within it is wrong.

All metrics are computed under four settings by toggling two conditions: with/without case ending

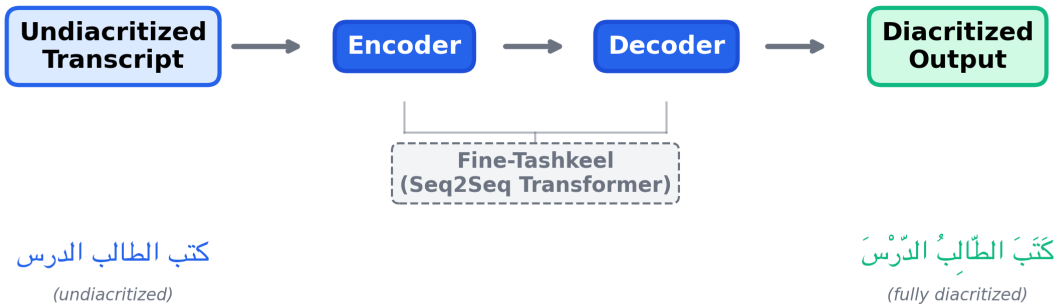

Figure 1: Overview of the Fine-Tashkeel Seq2Seq pipeline: undiacritized text is fed to the encoder; the decoder generates the fully diacritized sequence character by character.

(i'rab), and including/excluding positions with no diacritic (detailed four-setting results in Appendix 17). The official leaderboard uses the "Including no diacritic, With case ending" (WCE) setting.

6. Results

6.1. Official Leaderboard

Table 2 presents the official test leaderboard. Our Fine-Tashkeel system ranks 5th out of 7 participating teams with a DER of 10.56% and WER of 34.47%, 3.69 percentage points behind the top-ranked team. Notably, despite using only zero-shot inference without any task-specific training, our system outperforms teams 6 and 7 by a wide margin.

#	Team	DER↓	WER↓	SER↓
1	meshal	6.87	23.26	66.16
2	nadaadelmousa	7.04	24.39	71.65
3	naif_alharthi	7.51	25.34	73.48
4	nahian_abu	8.23	30.37	80.79
5	**Hassan (Ours)**	**10.56**	**34.47**	**79.88**
6	omarnj10	27.94	44.05	98.78
7	astral_fate	31.67	84.50	99.70

Table 2: Official test leaderboard (WCE, Including no diacritic). All metrics are percentages (%); lower is better. Our system (Fine-Tashkeel) ranks 5th.

6.2. Internal Model Comparison

Table 3 presents the top-performing models from our 18-configuration evaluation on the blind test set (Appendix 15). Fine-Tashkeel achieves the best DER (10.56%), followed closely by our ASR-Tashkeel pipeline (10.70%). Development set results are reported in Appendix 4.

Model	Method	DER	WER	SER
Fine-Tashkeel	**S2S / I**	**10.56**	**34.47**	**79.88**
ASR-Tashkeel	S2S / P	10.70	–	–
Shakkelha	T / I	13.14	39.47	83.84
Shakkala	T / I	19.70	56.37	96.95
CATT	T / I	28.34	48.17	100.0
Tashkeel-700M	T / FT	34.04	57.88	99.09

Table 3: Top 6 configurations on the blind test set, ranked by DER (%). *S2S* = Seq2Seq, *T* = text-only, *I* = zero-shot, *FT* = fine-tuned, *P* = ASR pipeline. Full results in Appendix 2.

6.3. Text-only vs. Multimodal Comparison

A key finding is the stark contrast between text-only and multimodal (ASR+Text) models. Table 4 summarizes this comparison (Appendix 14).

Category	Avg. DER	Best DER
Text-only (Inference)	44.15	10.56
ASR+Text (Inference)	70.57	42.82
Text-only (Fine-tuned)	44.10	34.04
ASR+Text (Fine-tuned)	54.61	54.61

Table 4: Average and best blind test DER (%) by model category. Text-only models outperform ASR+Text approaches across both settings.

Despite the task involving speech dictation, multimodal models such as Seamless M4T (DER 43.01%), Whisper variants (DER 84.71–97.36%), and ArTST (DER 42.82%) underperformed significantly. **Importantly, this result should be**

interpreted with care: these off-the-shelf ASR models are optimized for transcription rather than character-level diacritic generation, and their decoders default to unvoweled output distributions. The underperformance reflects a *task mismatch*, not an inherent limitation of audio information—purpose-built multimodal architectures with task-specific acoustic adaptation could benefit from acoustic cues to disambiguate homographs.

6.4. Comparison with Organizer Baselines

Table 6 (Appendix 3) presents the organizer's detailed baseline results on the test set across all four evaluation settings. Our inference-only DER of 10.56% (WCE, Incl. 0) is competitive with the organizer's fine-tuned Text+ASR baseline (9.91%), despite requiring no task-specific training.

6.5. Error Analysis

A per-nationality breakdown on the development set (Table 8, Appendix 5; visualized in Appendix 11) reveals significant dialect-dependent variation. Egyptian Arabic achieves the lowest DER (3.70%), benefiting from dominant training representation (1,026/2,327 samples), while Gulf dialects (Qatar 8.74%, Saudi 10.49%, Bahrain 11.17%) and Maghrebi dialects (Algeria 13.73%) show substantially higher error rates. Per-sentence analysis (Appendix 6) shows 33.1% of sentences achieve perfect diacritization, while only 3.1% exceed 30% DER—predominantly dialectal utterances (examples in Appendix 7). A diagnostic analysis of the baseline model (Appendix 8) reveals systematic errors in sukūn, damma, and tanwīn prediction, confirming that case endings and vowel ambiguity are the primary bottlenecks.

7. Discussion

Our inference-only DER of 10.56% is competitive with the organizer's fine-tuned baseline (9.91%), validating the Seq2Seq paradigm without task-specific training. A dev-test divergence analysis (Appendix 13) confirms Fine-Tashkeel as the most stable model (dev-to-test gap of only 2.46 pp). We acknowledge that our final submission relies on a text-only pipeline, bypassing the acoustic modality the shared task was designed to evaluate. Fine-tuning attempts encountered convergence issues (Appendix 18); off-the-shelf multimodal models proved poorly suited for character-level diacritization. As demonstrated by Shatnawi et al. (2024), effective speech-aware diacritization requires purpose-built architectures jointly modeling acoustic and textual features.

The 3.69 pp gap from the top-ranked system suggests avenues for improvement: (1) task-specific fine-tuning with larger corpora; (2) purpose-built multimodal architectures fusing acoustic embeddings with text representations; and (3) data augmentation targeting underrepresented dialects (Appendix 5).

8. Conclusion

We presented Fine-Tashkeel for the KSAA-2026 Shared Task, ranking 5th of 7 teams (DER 10.56%, WER 34.47%). Through evaluation of 18 models, we demonstrate that: (1) character-level Seq2Seq models outperform off-the-shelf multimodal and classifier-based approaches, reflecting task mismatch rather than an inherent limitation of acoustic information; (2) text-only inference achieves performance competitive with fine-tuned baselines; and (3) dialectal variation is the dominant error source (Egyptian 3.70% vs. Algerian 13.73% DER).

Limitations

Our system has limitations. First, as a text-only approach, it cannot leverage acoustic cues that could disambiguate homographs. Second, performance degrades on underrepresented dialects (Algerian, Bahraini). Third, we did not successfully fine-tune the Fine-Tashkeel model on the shared task data due to convergence issues, relying instead on zero-shot inference. Fourth, our evaluation is limited to the relatively small VoiceWall dataset ($\sim$5 hours), and generalization to other domains or larger corpora remains untested.

Ethics Statement

This work supports the development of Arabic language technology, specifically diacritization of speech dictation data. The shared task data were collected via the King Salman Global Academy's VoiceWall platform; consent and curation are the task organizers' responsibility, and we use the data as provided. Automated diacritization may contain errors; we recommend human validation before use in educational, religious, or accessibility applications. Our code supports reproducibility.

Acknowledgements

We thank the King Salman Global Academy for Arabic Language for providing the VoiceWall dataset and organizing the shared task (Al Wazrah et al., 2026). We also acknowledge the baseline implementations provided by Shatnawi et al. (2024).

References

Bashar Al-Rfooh, Gheith Abandah, and Rami Al-Rfou. 2023. Fine-tashkeel: Finetuning byte-level models for accurate arabic text diacritization. ArXiv preprint arXiv:2303.14588.

Asma Al Wazrah, Waad Alshammari, Rawan Almatham, Raghad Al-Rasheed, Afrah Altamimi, Rufael Marew, Sawsan Alqahtani, Hanan Aldarmaki, Abdullah Alharbi, Abdulrahman Alshehri, Mohamed Assar, Amal Almazrua, and Abdulrahman AlOsaimy. 2026. KSAA-2026 Shared Task on Arabic Speech Dictation with Automatic Diacritization. In *Proceedings of the 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7)*.

Faris Alasmary, Orjuwan Zaafarani, and Ahmad Ghannam. 2024. CATT: Character-based Arabic tashkeel transformer. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*, pages 250–257, Bangkok, Thailand. Association for Computational Linguistics.

Ahmad Barqawi and Taha Zerrouki. 2017. Shakkala: Arabic text vocalization. https://github.com/Barqawiz/Shakkala.

Bisher. 2025. qwen-1p5-diacritization. https://huggingface.co/Bisher/qwen-1p5-diacritization. Accessed: March 2026.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25:1–53.

Etherll. 2025. Tashkeel-700m model. https://huggingface.co/Etherll/Tashkeel-700M. Accessed: March 2026.

Ali Fadel, Ibraheem Tuffaha, Bara' Al-Jawarneh, and Mahmoud Al-Ayyoub. 2019. Neural Arabic text diacritization: State of the art results and a novel approach for machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 215–225, Hong Kong, China. Association for Computational Linguistics.

glonor. 2024. byt5-arabic-diacritization. https://huggingface.co/glonor/byt5-arabic-diacritization. Accessed: March 2026.

MaddoggProduction. 2026. whisper-l-v3-turbo-quran-lora-dataset-mix. https://huggingface.co/MaddoggProduction/whisper-l-v3-turbo-quran-lora-dataset-mix. Accessed: March 2026.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032. European Language Resources Association.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *PMLR*, pages 28492–28518.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual expressive and streaming speech translation. In *ArXiv preprint arXiv:2312.05187*.

Sara Shatnawi, Sawsan Alqahtani, and Hanan Aldarmaki. 2024. Automatic restoration of diacritics for speech data sets. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4166–4176. Association for Computational Linguistics.

tarteel-ai. 2022. whisper-base-ar-quran. https://huggingface.co/tarteel-

ai/whisper-base-ar-quran. Accessed: March 2026.

Hawau Toyin, Amirbek Djanibekov, Ajinkya Kulkarni, and Hanan Aldarmaki. 2023. ArTST: Arabic text and speech transformer. In *Proceedings of ArabicNLP 2023*, pages 41–51, Singapore (Hybrid). Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 6000–6010.

WajeehAzeemX. 2024. whisper-tiny-ar-tashkeel. https://huggingface.co/WajeehAzeemX/whisper-tiny-ar-tashkeel. Accessed: March 2026.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498. Association for Computational Linguistics.

Taha Zerrouki. 2020. *Towards An Open Platform For Arabic Language Processing*. Phd thesis, Ecole Nationale Supérieure d'informatique, Alger, Algérie.

1. Evaluation Metrics

System performance is evaluated using three complementary metrics. Let N_c, N_w, and N_s represent the total number of characters, words, and sentences in the reference text, respectively.

1. Diacritic Error Rate (DER): Measures the proportion of incorrectly predicted diacritics.

$$\text{DER} = \frac{S_c + D_c + I_c}{N_c} \tag{2}$$

2. Word Error Rate (WER): The primary metric. A word is marked incorrect if it contains at least one diacritic error, making it highly sensitive to case endings (*i'rab*).

$$\text{WER} = \frac{\text{Incorrect Words}}{N_w} \tag{3}$$

3. Sentence Error Rate (SER): A strict metric where a sentence is incorrect if any diacritic within it is flawed.

$$\text{SER} = \frac{\text{Incorrect Sentences}}{N_s} \tag{4}$$

These metrics are computed under four settings by toggling two binary conditions: (a) whether case endings (i'rab) are considered, and (b) whether character positions with no diacritic are included in the count. The official leaderboard uses "With case ending, Including no diacritic" (WCE, Incl. 0).

2. Full Internal Results

3. Organizer Baseline Results

4. Development Set Analysis

Fine-Tashkeel achieves the best dev DER (8.10%) among all models with reliable dev evaluation, followed by Shakkelha (10.26%) and Shakkala (11.64%). Fine-Tashkeel also shows the most consistent dev-to-test behavior, with a modest gap of only 2.46 percentage points (8.10% $\rightarrow$ 10.56%), indicating robust generalization.

5. Per-Nationality Detailed Analysis

Table 8 presents the full per-nationality breakdown of DER, WER, and SER on the development set, with and without case ending evaluation.

5.1. Gender Analysis

The model shows a modest gender gap: male speakers achieve 7.78% DER (WCE) compared to 8.62% for female speakers, a difference of 0.84 percentage points. This is likely due to the higher proportion of Egyptian male speakers in the dataset rather than an inherent gender bias in the model.

6. Error Distribution Analysis

Table 9 shows the distribution of per-sentence DER on the development set.

The average case ending impact on DER is -5.51%, meaning that removing case ending evaluation reduces DER by approximately 5.5 percentage points on average. This indicates that *i'rab* (grammatical case marking) remains a significant source of error, particularly for informal and dialectal speech where case endings are often phonologically reduced or absent.

7. Qualitative Error Examples

Table 10 presents representative error cases from the worst-performing predictions on the dev set. The errors fall into three main categories:

Model / Submission Name	Type	Method	DER (Dev)	DER (Blind) ↓	WER (Blind) ↓	SER (Blind) ↓
Inference (no task-specific training)						
Fine-Tashkeel (Al-Rfooh et al., 2023)	**Seq2Seq**	Inference	**8.10**	**10.56**	**34.47**	**79.88**
Shakkelha (Fadel et al., 2019)	Text-only	Inference	10.26	13.14	39.47	83.84
Shakkala (Barqawi and Zerrouki, 2017)	Text-only	Inference	11.64	19.70	56.37	96.95
CATT (Alasmary et al., 2024)	Text-only	Inference	–	28.34	48.17	100.00
FLAN-T5 (Chung et al., 2024)	Text-only	Inference	24.37	37.80	59.64	100.00
ArTST (Toyin et al., 2023)	ASR+Text	Inference	–	42.82	65.94	85.06
Seamless M4T (Seamless Communication et al., 2023)	ASR+Text	Inference	–	43.01	56.57	100.00
CAMeL-MLE (Obeid et al., 2020)	Text-only	Inference	28.71	45.89	82.39	100.00
ByT5 Glonor (glonor, 2024)	Text-only	Inference	15.28	46.96	69.64	100.00
Qwen-1.5 (Bisher, 2025)	Text-only	Inference	33.75	62.56	75.75	97.87
Mishkal (Zerrouki, 2020)	Text-only	Inference	17.37	76.76	90.49	99.39
Whisper (large-v3) (Radford et al., 2023)	ASR+Text	Inference	–	84.71	99.60	100.00
Tarteel Whisper (tarteel-ai, 2022)	ASR+Text	Inference	–	84.93	85.08	100.00
Whisper Quran LoRA (MaddoggProduction, 2026)	ASR+Text	Inference	–	97.36	99.98	100.00
mT5-base (Xue et al., 2021)	Text-only	Inference	82.17	99.74	99.97	100.00
Fine-tuned on task data						
Tashkeel-700M (Etherll, 2025)	Text-only	Fine-tune	22.08	34.04	57.88	99.09
Whisper Tiny Ar Tashkeel (WajeehAzeemX, 2024)	ASR+Text	Fine-tune	46.05	54.61	67.68	99.39
ByT5 Glonor (glonor, 2024)	Text-only	Fine-tune	23.12	54.15	77.57	100.00

Table 5: Comprehensive results for all 18 models evaluated. Models are grouped by method (inference vs. fine-tune) and sorted by blind test DER within each group. Dev scores computed using our evaluation script; blind test scores from the competition platform. "–" indicates dev scores excluded due to an evaluation library artifact that post-processed predictions. All values are percentages (%).

Evaluation Setting (%)	Text+ASR			Text-only			Fine-Tuned Text+ASR		
	DER	WER	SER	DER	WER	SER	DER	WER	SER
Including no diacritic									
With case ending	13.50	40.24	82.32	17.66	49.85	91.77	9.91	31.84	82.93
Without case ending	10.58	27.95	71.95	13.23	32.24	82.62	7.89	20.99	67.07
Excluding no diacritic									
With case ending	14.26	33.03	75.61	20.08	46.20	91.77	8.52	24.73	78.66
Without case ending	9.96	19.71	60.37	13.93	27.07	81.71	4.82	10.89	50.61

Table 6: Detailed organizer baseline results on the **test set**. Columns show DER/WER/SER for each configuration. Our inference-only DER (10.56%, WCE, Incl. 0) is competitive with the organizer's fine-tuned Text+ASR baseline (9.91%).

1. **Dialectal Mismatch:** The model applies MSA diacritization patterns to dialectal forms (e.g., Gazan عاوزه *'āwzah* 'she wants' receives MSA voweling).

2. **Morphological Ambiguity:** Words with multiple valid readings receive the wrong diacritic pattern (e.g., مجريات 'mujariyat' vs. 'mujriyat').

3. **Transliteration Errors:** Foreign terms and proper nouns are incorrectly diacritized due to absence from training data.

8. Baseline Diagnostic Analysis

To better understand the challenges of diacritizing speech-derived text, we conduct a diagnostic analysis using the official text-only baseline model on the development set. The baseline is a Transformer-based diacritization model pretrained on the Tashkeela corpus, operating purely on undiacritized text without access to acoustic information. While such models perform well on written Modern Standard Arabic (MSA), we observe substantial degradation when applied to speech transcripts in the KSAA dataset.

Sentence-Level Errors Using a strict Sentence Error Rate (SER) criterion—where a sentence is considered incorrect if any diacritic differs from the reference—we find that **95.0%** of sentences (247/260) contain at least one error. This reflects the inherently strict nature of full diacritization: even a single incorrect diacritic leads to a full sentence error. Nevertheless, the magnitude of this result highlights the difficulty of the task on speech-derived inputs.

Case Ending Errors Case endings (*i'rāb*) emerge as a dominant source of error. By com-

Rank	Model	DER (WCE)	WER (WCE)	SER (WCE)	DER (w/o CE)	WER (w/o CE)
1	**Fine-Tashkeel**	**8.10**	**20.46**	**67.31**	**13.86**	**43.66**
2	Shakkelha	10.26	24.37	72.31	15.92	47.03
3	Shakkala	11.64	27.89	78.85	17.35	50.02
4	ByT5 Glonor	15.28	38.43	94.23	20.59	56.84

Table 7: Models with reliable dev set scores, ranked by DER (WCE, Incl. 0). Several other models (ArTST, CATT, Seamless M4T, Ensemble, Postprocess) were excluded because their dev scores were artifacts of an evaluation library that inadvertently post-processed the predictions.

Nationality	N	DER (WCE)	WER (WCE)	SER (WCE)	DER (w/o CE)	WER (w/o CE)
Egypt	114	3.70	9.01	42.98	4.88	15.62
Palestine	8	4.71	12.66	75.00	14.67	55.70
Kuwait	12	5.75	14.86	50.00	15.49	53.38
Syria	1	5.65	12.50	100.00	16.03	62.50
Qatar	24	8.74	20.45	87.50	19.63	58.47
Saudi	41	10.49	28.67	90.24	18.17	60.37
Sudan	19	10.86	30.08	89.47	17.20	57.32
Bahrain	13	11.17	34.87	92.31	18.98	61.18
Algeria	28	13.73	35.05	92.86	20.79	62.70
Overall	**260**	**8.10**	**20.46**	**67.31**	**13.86**	**43.66**

Table 8: Full per-nationality DER/WER/SER breakdown on the dev set for Fine-Tashkeel, with and without case ending evaluation. The gap between best (Egypt, 3.70% DER) and worst (Algeria, 13.73% DER) nationality is 10.03 percentage points.

DER Range	Count	Percentage
0% (Perfect)	86	33.1%
0–5%	41	15.8%
5–10%	55	21.1%
10–15%	32	12.3%
15–20%	17	6.5%
20–30%	21	8.1%
30–50%	8	3.1%
Total	**260**	**100%**

Table 9: Distribution of per-sentence DER (WCE, Incl. 0) on the dev set. One-third of sentences are perfectly diacritized; the tail ($>$30% DER) consists exclusively of dialectal utterances.

paring the final diacritic of each sentence, we observe that **55.4%** of sentences exhibit incorrect case endings. This confirms that syntactically governed diacritics are particularly challenging to recover from text alone, especially in conversational and dialectal speech where such endings are often reduced, omitted, or ambiguous.

Diacritic Distribution Shift We compare the distribution of predicted diacritics against the ground truth. Three systematic failure modes emerge:

- **Sukūn underprediction** (-26.7%): difficulty modeling consonant closure, leading to 605 missed zero-vowel markers—the highest absolute error count.
- **Damma overprediction** ($+26.3\%$): confusion between vowel classes in ambiguous phonological contexts, contributing 231 errors.
- **Tanwīn overprediction**: severe systematic bias across all tanwīn types. Dammatan $+166.7\%$ (40 errors), Kasratan $+54.1\%$ (46 errors), Fathatan $+43.7\%$ (31 errors), indicating difficulty distinguishing indefinite case marking from standard vowel assignments.

High-frequency diacritics show contrasting patterns: *fatḥa* exhibits minor deviations (3.3%), while sukūn shows substantial errors (26.7%), indicating that errors concentrate in phonetically subtle (consonant closure) and grammatically complex (case marking) categories rather than simply low-frequency ones.

These results complement the modality comparison in Section 6: text-only models struggle with vowel ambiguity and case endings due to missing phonetic cues, while off-the-shelf multimodal systems fail to compensate due to task mismatch and error propagation (additional details in Appendix 9).

ID	DER	Reference	Prediction
Gaza_025	28.6	مَحْرَقَةُ رَفَحْ اَلْعَدُوّ يَنْتَقِمْ مِنَ اَلْآمِنِينْ *maḥraqatu rafaḥ al-ʿaduww yantaqim min al-āminīn* 'The massacre of Rafah; the enemy takes revenge on the safe ones'	مُحْرِقَةُ رِفْحِ الْعَدُوّ يَنْتَقِمُ مِنْ الْآمِنِينَ *muḥriqatu rifḥi al-ʿaduww yantaqimu min al-āmināna*
Qatar_422	32.5	إِتْسَلِّمْ عَلَى رَفِيجِكْ تَاخُذْ تِكْرِمْ جَارَكْ *itsallim ʿalā rafījik tākhidh tikrim jārak* 'Greet your friend, take [gifts], honor your neighbor'	إِتْسَلِمْ عَلَى رَفِيجِكَ تَاخُذْ تُكْرِمْ جَارَكَ *itsalim ʿalā rafījika tākhudh tukrim jāraka*
Sudan_121	32.6	وَاِمْتِحَنَّا وِاِجْتَزْنَا اَلْاِمْتِحَانَاتْ *wa-imtiḥannā wa-ijtaznā al-imtiḥānāt* 'We were tested and we passed the exams'	وَامْتَحَنَّا وَاجْتَزْنَا الامْتِحَانَات *wa-mtaḥannā wa-jtaznā al-imtiḥānāti*
Egypt_287	33.3	ثَنَا إِسْمَاعِيلُ بْنُ عَيَّاشٍ *thanā ismāʿīlu bnu ʿayyāshin* 'Isma'il ibn Ayyash narrated to us'	ثنا إِسْمَاعِيلُ بْنُ عَيَّاشٍ *thanā ismāʿīlu bnu ʿayyāshin*

Table 10: Representative worst-case error examples from the dev set. DER is per-sentence (%, WCE). Each cell shows the Arabic text, its transliteration, and (for reference) an English translation. Errors predominantly occur on dialectal forms and morphologically ambiguous words.

9. Text-only vs. Multimodal: Additional Details

The main text (Section 6) presents our modality comparison and its interpretation. Here we provide additional context on individual multimodal model failures.

Whisper-based models fine-tuned on Quranic Arabic (Tarteel Whisper, DER 84.93%; Whisper Quran LoRA, DER 97.36%) exhibit the most severe domain mismatch: formal tajweed diacritization rules are incompatible with conversational dialectal speech. The standard Whisper large-v3 model (DER 84.71%) produces predominantly unvoweled transcriptions, confirming that its decoder distribution is optimized for transcription rather than diacritization. Seamless M4T (DER 43.01%) and ArTST (DER 42.82%) perform better but still far below text-only approaches, as their architectures are not designed for character-level diacritic generation.

10. Sequence-to-Sequence Configuration

The `Fine-Tashkeel` model was configured with the following generation parameters to ensure deterministic and accurate character rendering:

- **Decoding Strategy:** Greedy Search (`do_sample = False`) to prioritize the most mathematically probable morphological sequence and prevent hallucinated characters.
- **Max Length:** 1024 tokens, sufficient to cover utterances in the dataset.
- **Penalty Formulation:** Length penalty applied to ensure the model does not prematurely truncate final case endings (*i'rab*).

11. Per-Nationality Visualization

Figure 2 visualizes the per-nationality DER and WER performance of Fine-Tashkeel on the development set. Nationalities are sorted by increasing DER and color-coded by dialect region: Nile Valley (Egypt, Sudan), Levantine (Palestine, Syria), Gulf (Kuwait, Qatar, Saudi, Bahrain), and Maghrebi (Algeria). The strong correlation between training data representation and model performance is evident, with Egypt ($n = 114$) achieving the lowest error rates and Algeria ($n = 28$) the highest.

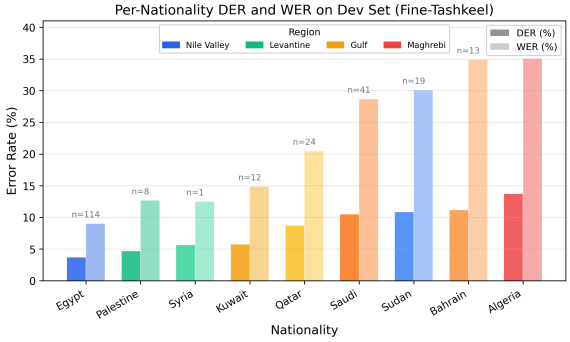

Figure 2: Per-nationality DER and WER (%) on the dev set, color-coded by dialect region. Sample counts are shown above each bar.

12. DER Distribution Visualization

Figure 3 presents the distribution of per-sentence DER values across the 260 development set samples. The distribution is heavily right-skewed: 33.1% of sentences achieve perfect diacritization (DER = 0%), while only 3.1% exceed 30% DER. This bimodal pattern suggests the model handles MSA-like and frequent dialectal patterns well, but struggles with rare dialectal forms.

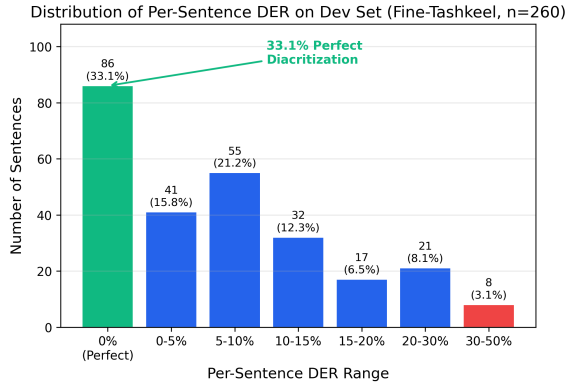

Figure 3: Distribution of per-sentence DER on the dev set ($n = 260$). One-third of sentences are perfectly diacritized; the tail consists exclusively of dialectal utterances.

13. Dev-Test Divergence Analysis

Figure 4 plots each model's dev DER (x-axis) against its blind test DER (y-axis) for models with reliable dev evaluation. Points on the diagonal indicate consistent performance; points far above indicate overfitting or distribution shift.

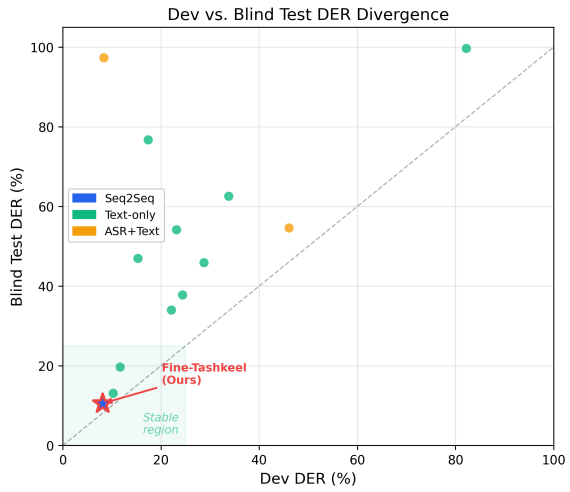

Figure 4: Dev vs. blind test DER for models with reliable dev evaluation. Fine-Tashkeel (star) shows the most stable generalization behavior.

Fine-Tashkeel is the most stable model among those with reliable dev evaluation, with a modest dev-to-test increase of only 2.46 percentage points (8.10% $\rightarrow$ 10.56%). Models with larger dev-to-test gaps (e.g., Shakkala: +8.06 pp, ByT5 Glonor: +31.68 pp) suggest sensitivity to dialectal distribution shift between the dev and test splits. Table 11 quantifies these divergences.

Model	Dev DER	Test DER	Δ
Fine-Tashkeel	8.10	10.56	**+2.46**
Shakkelha	10.26	13.14	+2.88
Shakkala	11.64	19.70	+8.06
ByT5 Glonor	15.28	46.96	+31.68
Mishkal	17.37	76.76	+59.39
FLAN-T5	24.37	37.80	+13.43
Qwen-1.5	33.75	62.56	+28.81

Table 11: Dev-to-test DER divergence (Δ) for models with reliable dev evaluation. Fine-Tashkeel shows the smallest gap, indicating robust generalization.

14. Text-only vs. Multimodal Visualization

Figure 5 visualizes the modality comparison presented in the main text (Table 4).

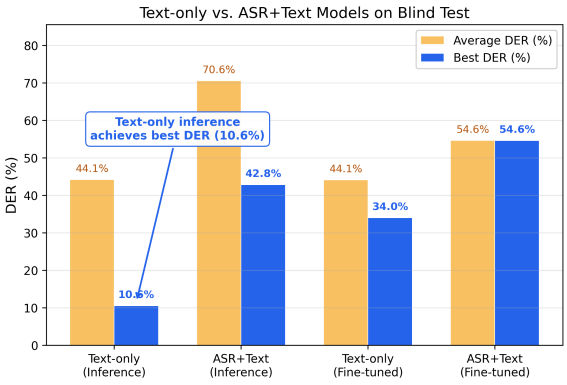

Figure 5: Average and best blind test DER by model category. Text-only inference models outperform ASR+Text approaches across all settings.

As discussed in the main text, this result reflects the task mismatch of available off-the-shelf models rather than an inherent limitation of incorporating acoustic information.

15. Full Model Comparison Visualization

Figure 6 provides a visual overview of all 18 evaluated configurations ranked by blind test DER. Models are color-coded by type (Seq2Seq, Text-only,

ASR+Text) and shading distinguishes inference from fine-tuned models.

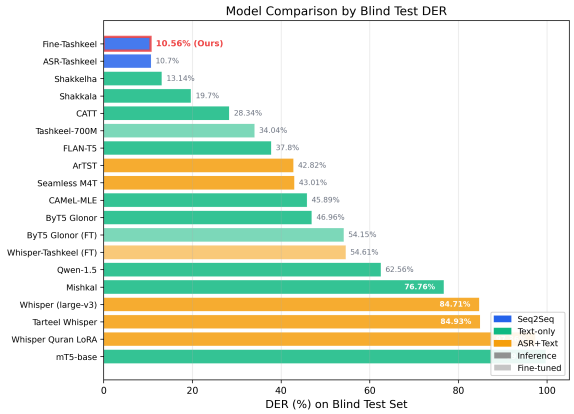

Figure 6: All evaluated models ranked by blind test DER. Fine-Tashkeel (highlighted) achieves the best DER among all configurations tested.

The visualization clearly shows the performance clustering: the top three models (Fine-Tashkeel, ASR-Tashkeel, Shakkelha) all fall below 15% DER and are all text-based Seq2Seq or text-only systems. The long tail of high-DER models is dominated by ASR+Text approaches, confirming that off-the-shelf multimodal models are poorly suited for this task.

16. ASR-Tashkeel Pipeline

Our second-best model (DER 10.70%) uses a two-stage ASR-Tashkeel pipeline that combines Whisper-based transcription with Fine-Tashkeel diacritization:

1. **Stage 1: ASR Transcription.** The input audio is processed by Whisper (large-v3) (Radford et al., 2023) to generate an undiacritized Arabic transcript. We use the standard inference configuration with Arabic language detection.

2. **Stage 2: Seq2Seq Diacritization.** The ASR-generated transcript is passed to Fine-Tashkeel (Al-Rfooh et al., 2023), which adds full diacritization using the same configuration described in Appendix 10.

This pipeline approach achieves a blind test DER of 10.70%, only 0.14 percentage points behind the text-only Fine-Tashkeel model (10.56%). The near-identical performance is attributable to Whisper's high-quality Arabic transcription: since the pipeline's diacritization stage receives a clean transcript, the quality bottleneck shifts to the diacritization model itself rather than transcription errors. This finding validates the use of ASR as a preprocessing stage when the ground-truth transcript is unavailable, while confirming that the clean transcript pathway remains marginally superior.

17. Detailed Four-Setting Results

Tables 12 and 13 present Fine-Tashkeel's results across all four evaluation settings on the development and test sets respectively, providing insight into the relative contributions of case ending errors and no-diacritic positions.

Setting	DER	WER	SER
Including no diacritic			
With case ending	8.10	20.46	67.31
Without case ending	13.86	43.66	83.46
Excluding no diacritic			
With case ending	7.43	14.90	58.46
Without case ending	15.13	39.71	77.31

Table 12: Fine-Tashkeel **dev set** performance across all four evaluation settings.

Setting	DER	WER	SER
Including no diacritic			
With case ending	10.56	34.47	79.88
Without case ending	7.46	20.81	68.29
Excluding no diacritic			
With case ending	11.56	29.45	76.83
Without case ending	6.86	14.32	64.33

Table 13: Fine-Tashkeel **test set** performance across all four evaluation settings (official CodaLab scores). Removing case endings reduces WER from 34.47% to 20.81% (−13.66 pp), confirming *i'rāb* as the primary error source.

The results reveal two important patterns on the dev set:

- **Case ending effect:** Evaluating without case ending (i'rab) increases DER from 8.10% to 13.86% (+5.76 pp). This counterintuitive result occurs because the "without case ending" setting excludes word-final positions from evaluation, removing positions where the model often predicts correctly (since many word-final diacritics are straightforward). The remaining interior positions have a higher proportion of ambiguous diacritics.

- **No-diacritic effect:** Excluding positions with no diacritic (Excl. 0) reduces DER from 8.10% to 7.43% (−0.67 pp), indicating the model

handles consonant clusters and sukun predictions well. The most challenging positions are those requiring active diacritic selection.

On the test set (Table 13), the pattern is more intuitive: removing case endings *reduces* WER from 34.47% to 20.81% (-13.66 pp), confirming that case endings are the primary error source on unseen data. Under the most lenient setting (WOCE, Excl. 0), our system achieves 6.86% DER and 14.32% WER. The full four-setting comparison across all teams is provided in Table 15 (Appendix 19).

Table 14 compares the top 4 models across both primary settings on the development set.

Model	DER (WCE, Incl. 0)	WER (WCE, Incl. 0)	DER (w/o CE, Incl. 0)	WER (w/o CE, Incl. 0)
Fine-Tashkeel	8.10	20.46	13.86	43.66
Shakkelha	10.26	24.37	15.92	47.03
Shakkala	11.64	27.89	17.35	50.02
ByT5 Glonor	15.28	38.43	20.59	56.84

Table 14: Top 4 models with reliable dev scores across two evaluation settings (WCE and without case ending, both Incl. 0).

18. Training Strategy Exploration

In addition to our primary zero-shot inference approach, we explored several training and adaptation strategies during development. None surpassed the zero-shot Fine-Tashkeel baseline on the blind test set, but they provide useful context for our final system choice.

18.1. Fine-Tashkeel Fine-Tuning

We attempted to fine-tune the Fine-Tashkeel model on the shared task training data (2,327 samples) for 5 epochs with a learning rate of 2×10^{-5}, batch size 30, and cosine learning rate scheduler. Data augmentation was applied via word dropout and character noise injection (yielding 3,523 augmented samples). However, the model failed to converge, producing 100% DER on the test set. This is discussed further in Limitations.

18.2. Curriculum Learning (ByT5)

We implemented a 3-stage progressive curriculum training strategy using the ByT5-Glonor model (glonor, 2024): (1) training on short, high-frequency MSA phrases; (2) adding medium-length dialectal utterances; (3) full dataset training. This achieved a dev DER of 28.19%, underperforming the standard ByT5 fine-tuning (dev DER 23.12%) and far below the zero-shot Fine-Tashkeel baseline (8.10%).

18.3. LoRA Fine-Tuning (Tashkeel-700M)

We applied Low-Rank Adaptation (LoRA) to the Tashkeel-700M model (Etherll, 2025) with rank $r = 16$, $\alpha = 32$, and 4-bit quantization to reduce memory requirements. This yielded a dev DER of 27.59%, comparable to the curriculum learning approach but similarly unable to match the zero-shot baseline.

18.4. ASR-Tashkeel Pipeline

Our two-stage Whisper+Fine-Tashkeel pipeline (detailed in Appendix 16) achieved a blind test DER of 10.70%, our second-best result and only 0.14 pp behind the text-only approach.

18.5. Ensemble Voting

We implemented majority voting across 6 models (Fine-Tashkeel, Shakkelha, Shakkala, ByT5, ArTST, CATT) at the character level. The ensemble achieved a dev DER of 7.48%, slightly outperforming the best individual model on the dev set (Fine-Tashkeel, 8.10%), but showing no improvement on the blind test set. This suggests the models share correlated errors on difficult dialectal examples.

18.6. Post-Processing Rules

We implemented rule-based post-processing including shadda reordering (ensuring shadda precedes the vowel diacritic) and tanween validation (checking consonant compatibility). These rules had minimal to slightly negative impact on DER, indicating that the Fine-Tashkeel model already produces well-formed diacritic sequences.

19. Official Leaderboard: Four-Setting Results

Table 15 presents the complete official test results for all 7 participating teams across all four evaluation settings, as reported on the CodaLab leaderboard. This extends the main leaderboard (Table 2), which reports only the primary setting (WCE, Incl. 0).

A key insight from this table is the impact of case endings on our system's performance. When excluding case endings (WOCE), our WER drops from 34.47% to 20.81%—a reduction of 13.66 percentage points—confirming that *i'rab* is the dominant error source. Furthermore, under the most lenient setting (WOCE, Excl. 0), our WER of 14.32%

	WCE, Incl. 0			WOCE, Incl. 0			WCE, Excl. 0			WOCE, Excl. 0		
Team	**DER**	**WER**	**SER**	**DER**	**WER**	**SER**	**DER**	**WER**	**SER**	**DER**	**WER**	**SER**
meshal	6.87	23.26	66.16	5.97	17.05	53.96	6.64	18.64	64.33	4.80	10.48	50.91
nadaadelmousa	7.04	24.39	71.65	6.17	17.96	55.18	6.92	18.97	68.90	5.15	11.06	50.61
naif_alharthi	7.51	25.34	73.48	6.60	18.66	60.06	7.51	19.95	70.73	5.68	11.85	55.49
nahian_abu	8.23	30.37	80.79	6.76	20.84	64.02	8.46	24.33	74.39	5.87	13.43	55.18
Hassan (Ours)	**10.56**	**34.47**	**79.88**	**7.46**	**20.81**	**68.29**	**11.56**	**29.45**	**76.83**	**6.86**	**14.32**	**64.33**
omarnj10	27.94	44.05	98.78	27.68	36.77	97.87	25.38	38.43	98.48	25.27	31.07	97.56
astral_fate	31.67	84.50	99.70	19.86	50.45	99.39	41.30	86.33	99.70	25.03	49.07	99.39

Table 15: Official test leaderboard across all four evaluation settings. All values are percentages (%); lower is better. WCE/WOCE = with/without case ending; Incl./Excl. 0 = including/excluding positions with no diacritic. Our system's WER drops from 34.47% (WCE) to 20.81% (WOCE), indicating case endings are the primary error source.

is competitive with nahian_abu (13.43%), narrowing the gap to just 0.89 pp. This suggests that our system's core diacritization capability is strong, and targeted improvements in case ending prediction could yield substantial gains in the primary metric.