Islamic University Journal of Applied Sciences (IUJAS)
https://journals.iu.edu.sa/jesc

Volume VIII, Issue 1, July 2026, Pages 280-298

Enhancing Arabic Dialectical Speech Recognition with Large Language Models

Rehab Albeladi[1]

[1] Department of Computer Science, Taibah University, Saudi Arabia

rbeladi@taibahu.edu.sa

Abstract

Automatic Speech Recognition (ASR) for Arabic faces unique challenges due to the language's diglossic nature, where Modern Standard Arabic (MSA) coexists with diverse dialects that differ substantially at lexical, phonological, and morphological levels. In this study, we present a comparative evaluation of state-of-the-art ASR models, Whisper (medium and large) and Wav2Vec2, applied to five Arabic varieties: Egyptian, Hijazi, Khaliji, Najdi, and MSA. We design a two-phase experiment: in the first phase, 50 speech samples per dialect are evaluated under three settings, raw transcriptions, and outputs corrected by two large language models (GPT and ALLAM). In the second phase, the best-performing ASR–LLM combination is scaled to the full test set for each dialect. Performance is assessed using Word Error Rate (WER) and semantic similarity distance. Results show that Whisper-Large combined with GPT consistently outperforms other systems across dialects. Improvements are especially pronounced in dialectal speech, while MSA achieves the lowest baseline WER. These findings highlight both the persistent difficulty of dialectal Arabic ASR and the effectiveness of LLM-based post-processing for reducing recognition errors.

Keywords: Automatic Speech Recognition; Large Language Models; Arabic Dialects.

https://doi.org/10.63070/jesc.2026.028
Received 15 February 2026; Revised 20 April 2026; Accepted 24 April 2026.
Available online 30 April 2026.
Published by Islamic University of Madinah on behalf of *Islamic University Journal of Applied Sciences.*
This is an open access article published under the Creative Commons Attribution 4.0 International license (CC BY 4.0):https://creativecommons.org/licenses/by/4.0/

1. Introduction

Automatic speech recognition (ASR) systems have made substantial progress in recent years, driven by advances in self-supervised learning and large-scale pretraining. Models such as Whisper [1] and Wav2Vec2.0 [2] have achieved near-human performance on several benchmark datasets, particularly for high-resource languages like English. However, the application of ASR to Arabic speech—especially its dialectal varieties—remains a significant challenge. Arabic is characterized by a diglossic structure, with Modern Standard Arabic (MSA) used in formal settings and a diverse set of regional dialects used in everyday speech. These dialects differ considerably in phonology, morphology, syntax, and lexicon, and they are often underrepresented in ASR training corpora.

Despite increasing interest in Arabic speech technologies, few studies have systematically evaluated the performance of state-of-the-art ASR models across a broad set of Arabic dialects. Moreover, raw ASR outputs often lack fluency and readability, particularly when dealing with dialectal input. Recent developments in large language models (LLMs), such as GPT-4 and Arabic-specific models like ALLAM, offer promising tools for post-processing ASR outputs. These models can correct transcription errors, restore grammaticality, and preserve meaning through context-aware generation.

In this work, we investigate how well two leading ASR systems—Whisper and Wav2Vec2—perform on Arabic dialectal speech using the SADA dataset, a multi-dialectal corpus covering five major Arabic varieties. We further examine the potential of LLMs to improve ASR outputs through generative correction, comparing the performance of GPT and ALLAM in this task. Our evaluation employs both Word Error Rate (WER), a traditional metric for transcription accuracy, and Semantic Distance (SemDist), which quantifies the preservation of meaning between hypotheses and reference texts. By analyzing results across dialects and system configurations, we aim to offer insights into the strengths and limitations of current tools for Arabic ASR, and to highlight the role of LLMs in bridging the gap in low-resource or high-variance speech settings. The focus of this study is on quantitative evaluation through objective metrics such as WER and semantic distance, complemented by statistical analysis across dialects. A qualitative investigation of recognition or correction errors is beyond the scope of this work.

2. Literature Review

This section presents a review of the existing literature on Automatic Speech Recognition (ASR) for the Arabic language. It begins with an overview of the field, outlining its principal challenges and historical development. Subsequently, it examines the application of deep learning models to Arabic ASR. The section then discusses the emerging practice of integrating Large Language Models (LLMs)

into ASR pipelines for output refinement. Finally, it concludes with a summary of relevant comparative studies that have benchmarked the performance of various ASR systems, providing context for the present study.

2.1 Overview of Arabic ASR Research

The evolution of Automatic Speech Recognition for Arabic is characterized by efforts to address a set of persistent linguistic challenges. The most significant of these is the diglossic state of the language, which involves a distinction between Modern Standard Arabic (MSA)—the form used in formal, official, and written contexts—and numerous colloquial dialects used as a spoken language for daily communication [3]. These dialects are not merely variations in accent but constitute distinct linguistic systems with substantial lexical, phonological, morphological, and syntactic divergences from MSA. For instance, the interrogative phrase "how are you?" translates to kayfa ḥāluk in MSA, but is commonly rendered as izzayak in Egyptian Arabic, shlōnak in Gulf Arabic, and kīfāk in Levantine Arabic. This degree of lexical divergence makes ASR systems trained exclusively on MSA data fail to recognize basic conversational phrases from dialectal speech.

Initial research in Arabic ASR focused predominantly on MSA. This focus was a pragmatic consequence of the greater availability of structured, annotated datasets derived from formal sources, such as broadcast news corpora. The main technology of this period was the Hidden Markov Model (HMM) framework, typically utilizing Gaussian Mixture Models (GMMs) to model the acoustic states of phonetic units [4]. The application of these GMM-HMM systems to Arabic was complicated by the language's complex morphology. A single Arabic root can yield hundreds of word forms through the affixation of proclitics (e.g., و"and", ف"so") and enclitics (e.g., pronoun suffixes). This morphological complexity results in high out-of-vocabulary (OOV) rates, thereby limiting model performance [5].

The performance of these MSA-centric models on dialectal Arabic exhibited substantial degradation. This linguistic gap underscored the necessity for dialect-specific resources. Consequently, a significant focus of subsequent research has been the development of large-scale, annotated speech corpora for various dialects. The creation of resources such as the SADA (Saudi Audio Dataset for Arabic) for Saudi dialects [6] and the Arabic Multi-Genre Broadcast MGB-2 dataset [7] are important contributions. The development of such corpora is a resource-intensive process that involves recording, manual transcription, and careful annotation to capture the phenomena of spontaneous speech, including disfluencies and code-switching. These datasets now provide the necessary foundation for the development and robust evaluation of dialect-aware ASR systems.

2.2 Deep Learning-Based ASR for Arabic

The application of deep learning to ASR marked a significant methodological shift, leading to substantial performance improvements and addressing many of the limitations of GMM-HMM systems. The initial integration of deep learning in ASR involved replacing GMMs with Deep Neural Networks (DNNs) within the hybrid GMM-HMM framework, enabling more discriminative modeling of acoustic features and resulting in immediate performance gains [8].

Subsequently, Recurrent Neural Networks (RNNs) architectures, including Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, became prevalent. The capacity of these models to capture long-range temporal dependencies in the speech signal was well-suited for language processing. For Arabic, this translated to an improved ability to model coarticulation effects and contextual phonetic information across long utterances, mitigating a known weakness of the first-order Markov assumption in HMMs [9].

A more fundamental change was the adoption of end-to-end (E2E) models. In these architectures, a sequence of acoustic features is directly mapped to a sequence of characters or words, unifying the acoustic, pronunciation, and language model components of traditional ASR into a single, jointly trained neural network. Particularly beneficial for Arabic dialects, this approach eliminates the need for manually constructed phonetic dictionaries, which are often unavailable for unwritten colloquial variants. Two primary E2E frameworks became prominent: Connectionist Temporal Classification (CTC) [10] and attention-based sequence-to-sequence (seq2seq) [11] models. CTC-based models learn a monotonic alignment between acoustic frames and output labels, while attention mechanisms permit the model to learn dynamic, non-monotonic alignments.

Current state-of-the-art systems are largely based on large-scale, pre-trained models that utilize the transformer architecture. The methodology of pre-training on extensive datasets and subsequently fine-tuning on a specific task has proven to be highly effective. Wav2Vec 2.0 [2] is a prominent example of self-supervised learning in this domain. These models are fine-tuneable on smaller, labeled dialectal datasets, which for instance, demonstrated effectively for Egyptian Arabic [12].

In parallel, Whisper [1]—trained in a weakly-supervised fashion on around 680,000 hours of multilingual, multitask audio—illustrates the potential of large-scale models in delivering "zero-shot" ASR performance. It can transcribe languages and dialects without explicit supervised fine-tuning, offering a robust baseline for under-resourced Arabic dialects [13].

2.3 LLM Integration in ASR Pipelines

The integration of large language models (LLMs) into the post-processing pipeline of automatic speech recognition (ASR) systems has emerged as a promising approach for error correction [14]. Traditional ASR systems, while increasingly accurate, still produce errors due to factors such as acoustic ambiguity, speaker variability, and domain mismatch. LLMs, with their advanced contextual understanding and generative capabilities, offer a promising solution for refining ASR outputs by correcting transcription errors and enhancing overall readability and accuracy.

Recent research has demonstrated that LLMs can be employed in both zero-shot and few-shot settings to perform ASR error correction without the need for extensive retraining on ASR-specific data. These models can leverage N-best lists—multiple hypotheses generated by the ASR system—to provide richer context for correction, or operate directly on the 1-best output for more generalizable solutions [15]. Approaches such as generative error correction (GEC) utilize LLMs to generate improved transcriptions by either unconstrained generation or by selecting the most plausible candidate from the N-best list [16]. Another approach investigated in the literature is in-context learning and prompt engineering, in which examples or specific instructions are provided in the prompt, so that LLMs can be guided to perform more accurate corrections, even without explicit retraining [17].

While an experimental study showed that LLM was unable to correct transcription errors [18], recent research has demonstrated that LLM-based correction reduces word error rate (WER) and improves transcription quality across a variety of ASR architectures, including both transducer and attention-based encoder-decoder models [15][19][20]. Notably, LLMs have shown remarkable generalization capabilities, effectively correcting outputs from different ASR systems even in black-box scenarios where the underlying ASR model is unknown [15]. Multilingual LLMs have also proven capable of correcting ASR errors in a wide range of languages, leveraging their cross-lingual knowledge to provide robust performance even when only the 1-best hypothesis is available [19].

For Arabic ASR, the post-correction stage has considerable utility. LLMs can be employed to address a range of prevalent error types, including the confusion between phonetically similar but orthographically distinct letters, errors in grammatical agreement (e.g., subject-verb, noun-adjective), which are common in Arabic and are frequently missed by ASR systems that lack deep syntactic awareness. However, the application of LLMs for ASR post-processing and error correction in Arabic speech remains less explored compared to English. Emerging research has begun to leverage LLMs for code-switched Arabic-English ASR, utilizing models like GPT-4 to generate synthetic code-switched data and improve ASR performance through N-best list rescoring and language model interpolation [24].

2.4 Comparative ASR Studies

As the variety of ASR models and methodologies for Arabic has expanded, the need for rigorous, data-driven comparison has become critical for benchmarking performance and guiding future research [7]. These comparative studies are essential for quantifying the relative merits of different architectures and understanding the specific conditions under which certain models excel. Initial comparative works focused on benchmarking traditional GMM-HMM systems against the first wave of hybrid DNN-HMM and end-to-end RNN-based models. These studies consistently reported the superior performance of deep learning approaches, particularly when larger training datasets were utilized [22][23].

The current focus of comparative analysis is on the new generation of large pre-trained models. Researchers are actively benchmarking fine-tuned Wav2Vec models against Whisper's zero-shot and fine-tuned performance across various Arabic speech datasets, such as the Multi-Genre Broadcast (MGB) challenge datasets [13][24]. These studies often reveal a trade-off: Whisper's broad, weakly-supervised pre-training provides a strong generalist model, while a Wav2Vec 2.0 model fine-tuned on a high-quality, domain-specific dataset can often achieve lower error rates for that specific domain or dialect. Recent comprehensive benchmarking efforts like the Open ASR Leaderboard [25] have systematically evaluated models like Whisper against other state-of-the-art systems, providing valuable insights into their capabilities on both Modern Standard and Dialectal Arabic. These benchmarks highlight that while pre-trained models have significantly advanced the field, performance can still vary widely depending on the specific dialect and acoustic conditions [13].

A critical aspect of these comparative studies is the choice of evaluation metric. For decades, the predominant metric has been the Word Error Rate (WER), which algorithmically computes a distance based on word-level substitutions, insertions, and deletions. However, the linguistic community has long critiqued WER's shortcomings for morphologically rich languages like Arabic. The tendency of WER to over-penalize minor morphological errors that may not impede human comprehension is a known issue [26]. In Arabic, for example, consider two potential errors in transcribing a spoken utterance. An error that changes kitābuhum ("their book") to kitābuhunna ("their [feminine] book") is a minor disagreement in grammatical gender. An error that changes it to maktabuhum ("their office") is a major semantic failure. Both scenarios could register as a single word error under the WER metric, despite their vastly different impact on meaning. This discrepancy illustrates that WER can be a poor proxy for semantic accuracy in Arabic [27].

The inadequacy of WER has motivated the use of semantically-oriented evaluation metrics. This research will therefore also employ Semantic Distance (SemDist), a metric that uses sentence embeddings from models like Sentence-BERT [28] to quantify the semantic similarity between an ASR hypothesis and a ground truth reference. The use of sentence-level semantic similarity has been explored in various Arabic NLP tasks, demonstrating its utility in capturing meaning beyond lexical overlap [29]. By representing sentences as vectors in a semantic space, SemDist evaluates the preservation of meaning, providing a useful complement to WER.

3. Methodology

Let X denote a set of Arabic speech utterances, each associated with a dialect label $d \in D$, where D is the set of dialects represented in the dataset. The objective is to evaluate the performance of state-of-the-art ASR models on transcribing X into text, and to assess whether post-processing these transcriptions with large language models (LLMs) can further optimize transcription quality. The evaluation is conducted using both Word Error Rate (WER) and Semantic Distance (SemDist) as metrics.

Formally, for each utterance $x \in X$, the ASR model f_{ASR} produces a hypothesis $h = f_{ASR}(x)$. The LLM post-processing step is defined as $h' = f_{LLM}(h)$. Both h and the final output h' are compared to the ground truth transcription y using WER and SemDist.

The experimental pipeline (figure 1) consists of the following stages:

1. ASR Inference: Each audio sample is transcribed using the benchmarked ASR models.
2. LLM Post-Processing: The ASR-generated hypotheses are passed to LLMs for error correction and refinement.
3. Evaluation: The outputs are evaluated against reference transcriptions using WER and SemDist.

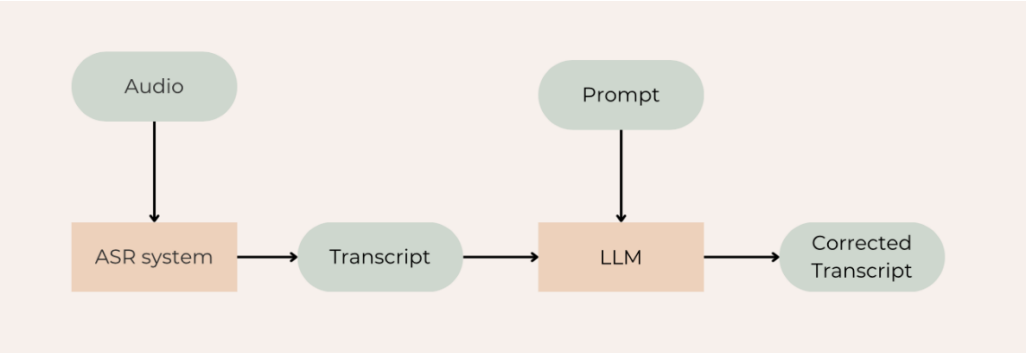


Fig. 1: Experimental Pipeline

This architecture enables a modular comparison of ASR models and the impact of LLM-based post-processing. Both the ASR models and LLMs are evaluated in a zero-shot setting, without any fine-tuning on the dataset or specific examples to guide the language model.

3.1 Dataset and pre-processing

The Saudi Audio Dataset for Arabic (SADA) [6] is a recently released benchmark that contains over 600 hours of audio covering multiple Arabic dialects. Each entry consists of an audio file, a ground-truth transcription, and metadata including dialect classification. All audio files were provided in WAV format with a sampling rate of 16 kHz and 16-bit PCM encoding. The dataset includes many Arabic dialects, mostly Saudi dialects, in addition to Egyptian, Yemeni, Levantine, besides Main Stream Arabic. The audio comes from television shows covering both spontaneous and speaking styles in diverse genres.

The dataset is split into training, validation, and test sets. For this study, we only use the test split of the SADA dataset. The test set contains 6193 audio segments. The average segment length in the set is 6.89 seconds while the median is 3.34. Figure 2 shows the speakers' genders and environment distribution, while Table 1 shows the dialects statistics of the test set in addition to calculated basic noise analytics. Out of 6193 entries in the test set, 134 contain "غير واضح" which means "unclear" in the transcription field. This means that these segments were not fully transcribed by the dataset providers. Such records are excluded from the evaluation set in the preprocessing step to guarantee that automatic speech recognition (ASR) performance metrics are computed only on utterances with fully verified reference transcriptions. To maintain consistency, transcripts are also normalized using Whisper's BasicTextNormalizer.

Table 1: SADA test set dialects

SpeakerDialect	*No. of audio segments*	*% of test set*	*Average SNR*	*Average clipping rate*
Najdi	1704	27.52%	13.32	0.0041
More than 1 speaker	1320	21.31%	13.58	0.0013
Khaliji	1150	18.57%	12.83	0.0034
Hijazi	809	13.06%	12.81	0.0035
Unknown	765	12.35%	11.71	0.0041
Notapplicable	167	2.70%	17.64	0.0042
ModernStandardArabic	157	2.54%	12.73	0.0017
Egyptian	96	1.55%	12.74	0.0034
Shamali	18	0.29%	15.61	0.0039
Yemini	7	0.11%	10.97	0.0036

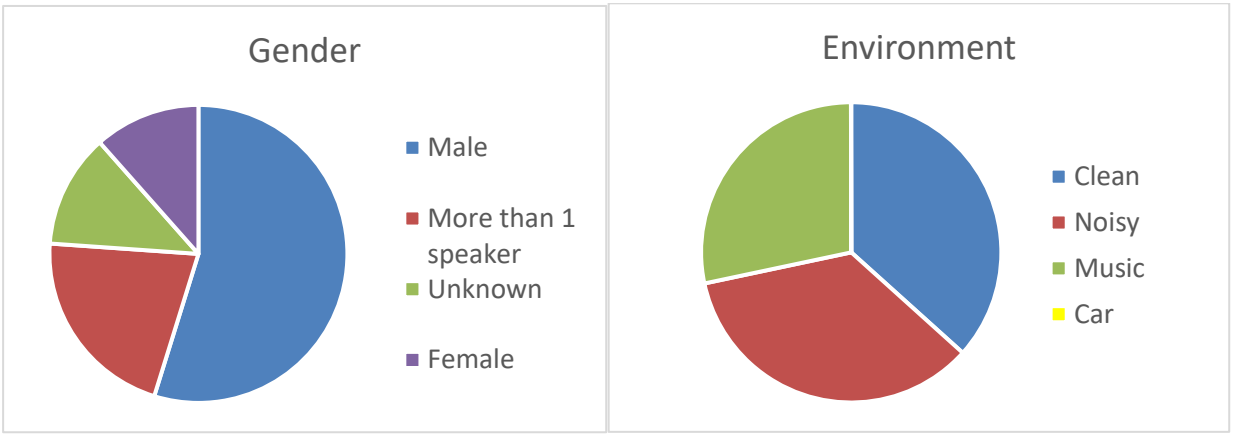


Fig. 2: Test set gender and environment distribution

3.2 ASR models

Two state-of-art ASR models are used in this study, namely, Whisper and Wav2vec. Whisper is an ASR system developed by OpenAI, designed to handle a wide range of languages and dialects. The architecture is based on an encoder-decoder Transformer, where input audio is first converted into a log-Mel spectrogram and then processed by the model to generate transcriptions. Whisper was trained on an extensive dataset exceeding 680,000 hours of multilingual and multitask supervised data, which includes a significant proportion of non-English audio. This large-scale, diverse training corpus enables Whisper to generalize robustly across languages, accents, and acoustic conditions, making it particularly effective for zero-shot transcription tasks [1]. Whisper is available in several pre-trained checkpoints of increasing size and capacity—tiny, base, small, medium, and large (including large-v2 and large-v3)—with each larger model offering improved transcription accuracy at the cost of greater computational requirements. Both the medium and large variants are applied in this study. Both models are multilingual and support direct transcription of Arabic speech, making them suitable for comparative studies involving dialectal Arabic.

Wav2Vec 2.0 is a self-supervised speech representation model developed by Facebook AI, which learns directly from raw audio waveforms. The jonatasgrosman/wav2vec2-large-xlsr-53-arabic model is a fine-tuned version of the large XLSR-53 checkpoint, specifically adapted for Arabic speech recognition. This model was trained using the Common Voice 6.1 and Arabic Speech Corpus datasets, allowing it to capture the phonetic and lexical characteristics of Arabic speech. The model operates at a 16kHz sampling rate. In its benchmark evaluations, the model achieved a WER of approximately 39.6% on the Common Voice Arabic test set.

3.3 LLMs

For the post processing of the transcribed audios, two large language models are employed. GPT (Generative Pre-trained Transformer) represents a family of large-scale language models developed by OpenAI, notable for their advanced natural language understanding and generation capabilities. GPT models are built on the Transformer architecture, leveraging self-attention mechanisms to capture complex contextual relationships within text. These models are pre-trained on vast and diverse text corpora, encompassing hundreds of billions of parameters, which enables them to generate highly coherent and contextually relevant responses to a wide range of prompts [30]. Recent versions, such as GPT-4, have further improved the model's ability to maintain context over longer interactions, handle multilingual content—including Arabic—and provide accurate outputs across a variety of applications. The model supports both zero-shot and in-context learning, making it appropriate for tasks such as ASR output correction.

The second model is ALLAM [31], which is a large language model specifically designed for Arabic language, addressing the unique linguistic characteristics of Arabic. Unlike general-purpose models such as GPT, ALLAM is trained predominantly on Arabic data, including Modern Standard Arabic and a wide spectrum of dialects, as well as English. Its architecture is based on the Transformer framework, similar to GPT, but with optimizations and pre-training strategies tailored for Arabic NLP tasks. The model demonstrates strong performance in a range of applications, including text generation, and summarization.

4. Experiments

4.1 Experiments setup

To evaluate the performance of automatic speech recognition (ASR) and language models on dialectal Arabic speech, we conducted a two-phase experimental setup. In the first phase, we compared the outputs of three ASR models—Whisper-medium, Whisper-large, and Wav2Vec 2.0—under three distinct conditions: (i) raw ASR output, (ii) ASR output corrected by GPT, and (iii) ASR output corrected by ALLAM. All models were evaluated in a zero-shot setting, without any fine-tuning or domain-specific adaptation. For each Arabic dialect under study (Khaliji, Najdi, Hijazi, Egyptian, and MSA), 50 audio segments were sampled from the test portion of the SADA dataset. These audio files were transcribed using each ASR model, and the resulting transcripts were either left unaltered (raw), or passed to an LLM for post-processing. Post-processing with GPT was carried out using GPT-4 via the OpenAI API (gpt-4 model), with carefully constructed prompts to encourage correction of standard

word forms while preserving dialectal meaning. For ALLAM, we used the ALLaM-7B model hosted on Hugging Face, accessed through the transformers library, and prompted similarly to correct the ASR output. This setup resulted in a total of nine system variants per dialect (three ASR models × three settings). To ensure fair comparison, evaluation metrics were computed for each system on normalized text.

In the second phase, the best-performing ASR and LLM pipeline from phase 1 was selected based on aggregate performance across dialects using two evaluation metrics: Word Error Rate (WER) and Semantic Distance (SemDist). The selected system was then applied to the entire test set of each dialect in the SADA corpus. This allowed us to conduct a more comprehensive evaluation of how well the optimal ASR+LLM configuration performs in transcribing dialectal Arabic speech and whether it generalizes effectively beyond the limited sample used in the first phase. WER was computed using standard Levenshtein distance (using python's jiwer library), while SemDist measured the semantic similarity between reference and hypothesis sentences using sentence embeddings and cosine distance.

4.2 Results and discussion

This section presents the evaluation results of ASR performance across five Arabic dialects using Whisper and Wav2Vec2 models, with and without post-processing by large language models (GPT and ALLAM). The results are analyzed in terms of Word Error Rate (WER) and Semantic Distance (SemDist) to assess both surface-level accuracy and semantic preservation.

Table 2: Phase 1 results

	WER					*SemDist*				
	Egy	Hjz	Klj	MSA	Njd	Egy	Hjz	Klj	MSA	Njd
Wav2Vec(raw)	0.96	0.93	0.92	0.88	0.94	0.39	0.42	0.40	0.46	0.35
Wav2Vec (GPT)	0.79	0.84	0.76	0.71	0.75	0.38	0.40	0.37	0.41	0.37
Wav2Vec (Allam)	0.88	0.91	0.98	0.84	0.89	0.43	0.43	0.46	0.48	0.41
Whisper-md (raw)	0.49	0.62	0.61	0.60	0.56	0.25	0.30	0.27	0.24	0.31
Whisper-md (GPT)	0.42	0.54	0.52	0.43	0.52	0.21	0.30	0.26	0.20	0.23
Whisper-md (Allam)	0.53	0.62	0.71	0.56	0.59	0.29	0.33	0.31	0.23	0.32
Whisper-lg (raw)	0.49	0.53	0.53	0.44	0.53	0.22	0.26	0.28	0.19	0.29
Whisper-lg (GPT)	**0.41**	**0.49**	**0.51**	**0.40**	**0.48**	**0.19**	**0.24**	**0.25**	**0.15**	**0.22**
Whisper-lg (Allam)	0.49	0.56	0.67	0.59	0.57	0.24	0.28	0.30	0.21	0.32

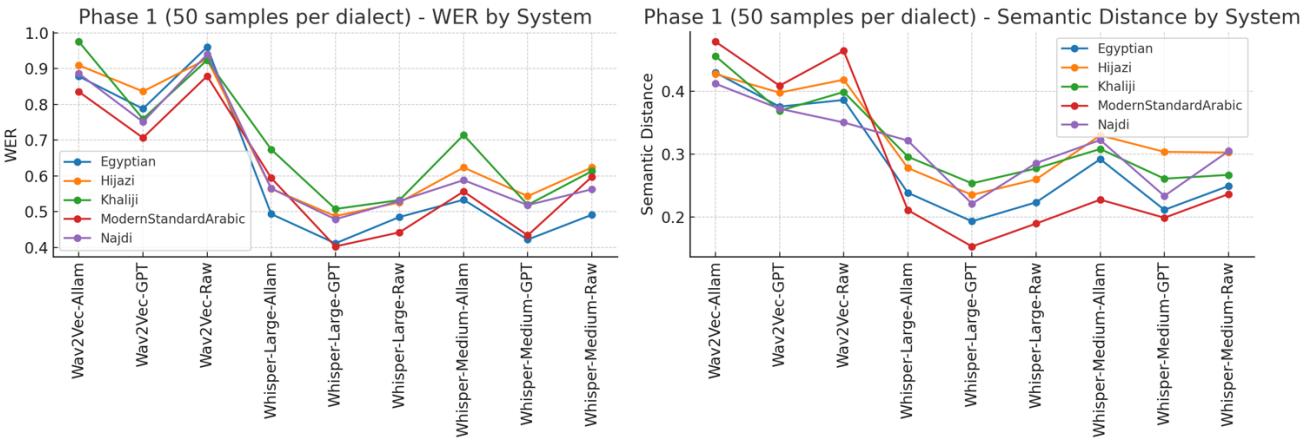


Fig. 3: WER and SemDist for phase 1

Results from the first phase are shown in table 2 and depicted in figure 3. They show that Whisper models consistently outperformed Wav2Vec2 across all dialects. In particular, Whisper-Large achieved the lowest Word Error Rate and Semantic Distance, outperforming Whisper-Medium and Wav2Vec2 in nearly all cases. This aligns with findings from Radford et al. [1], who demonstrated Whisper's robustness to noisy environments and multilingual input, partly due to its large-scale weakly supervised training on web-scale data. In contrast, Wav2Vec2, while strong in certain clean speech settings [2], appears less adapted to the pronunciation variability, loanwords, and informal constructions common in dialectal Arabic. The open universal Arabic ASR leaderboard [25] also shows that Whisper Large outperforms Whisper medium and Wav2Vec on SADA dataset.

Results also show that LLM post-processing appears to be consistently beneficial, with GPT showing more pronounced improvements than ALLAM. While ALLAM reduced WER in several cases, the gains were generally smaller than those achieved by GPT, and in a few instances, semantic distance was not improved. Notably, Wav2Vec2 benefited most from GPT correction, with large relative WER reductions. However, even after LLM correction, Wav2Vec2 lagged behind Whisper-based systems.

Dialectal variation also influenced results. Across all systems, MSA achieved the lowest error rates (best WER: 0.4031 with Whisper-Large-GPT), while other dialects such as Khaliji and Najdi consistently recorded higher WER values. This aligns with prior work [32] indicating that dialectal variability and phonological divergence from MSA in addition to resource scarcity introduce greater transcription challenges. Egyptian dialect also performed better than the other Saudi dialects. This may indicate underrepresentation in training data of the ASR systems.

The findings from Phase 1 strongly indicated Whisper-Large with GPT correction as the optimal configuration, offering the lowest WER and Semantic Distance results across all dialects. This informed the system selection for Phase 2.

Table 3: Phase 2 results

	Egyptian	*Hijazi*	*Khaliji*	*MSA*	*Najdi*
WER for Whisper-lg (raw)	0.59	0.53	0.56	0.32	0.49
WER for Whisper-lg (GPT-corrected)	0.52	0.49	0.51	0.35	0.48
SemDist for Whisper-lg (raw)	0.27	0.26	0.27	0.13	0.28
SemDist for Whisper-lg (GPT-corrected)	0.26	0.25	0.25	0.13	0.26

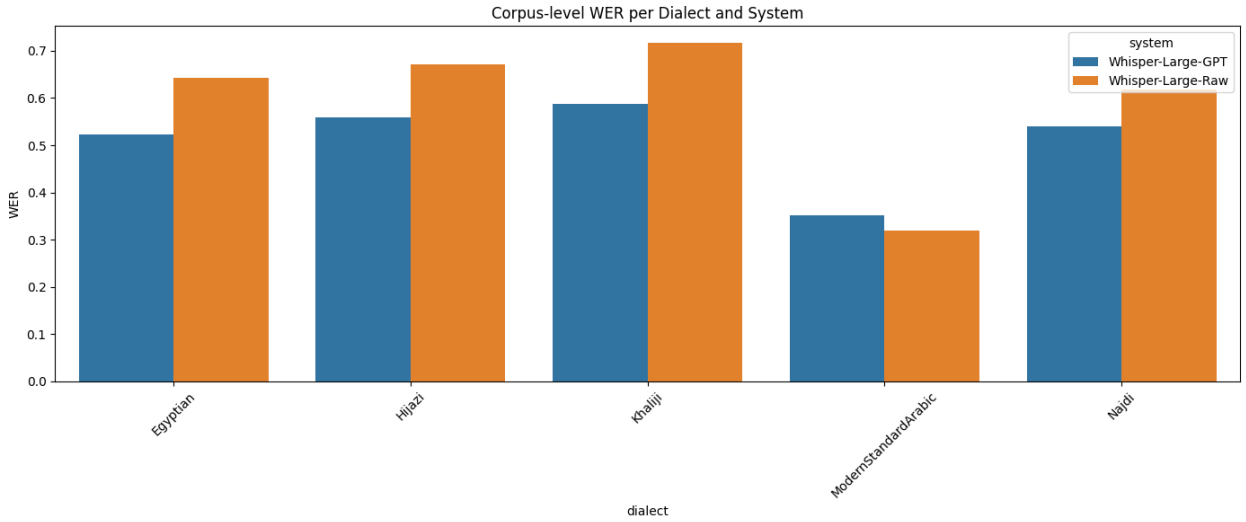


Fig. 4(a): WER for phase 2

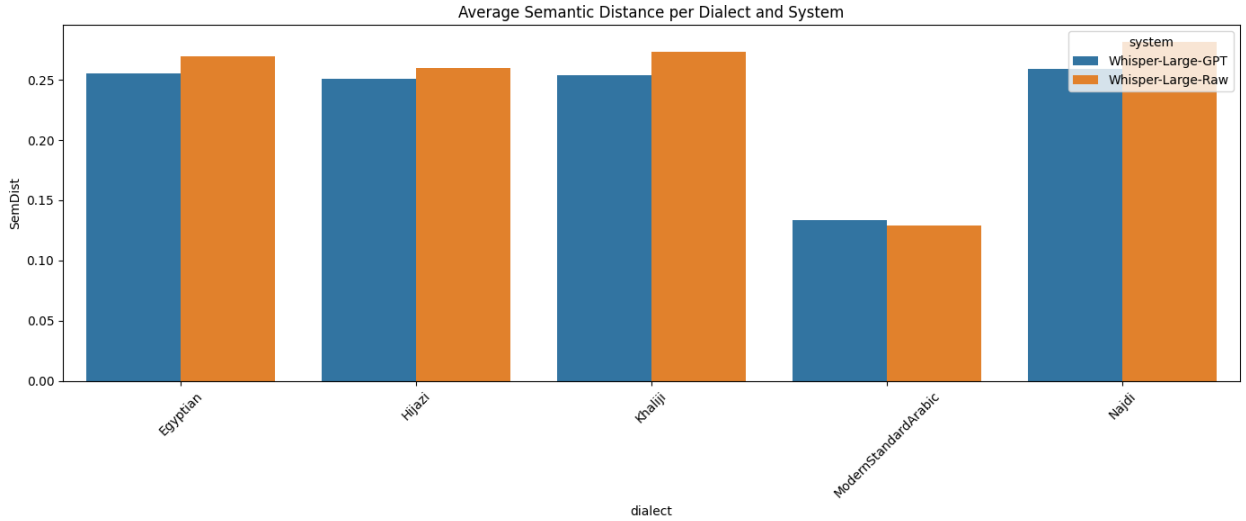


Fig. 4(b): SemDist for phase2

Phase 2 applied Whisper Large with GPT post-processing to the entire test set for each dialect, allowing for a more comprehensive performance evaluation, with Whisper-Large (raw) serving as a baseline. The initial results showed very high WER for some dialects datasets, and upon inspection, we found that a few audio segments were transcribed into repeated words which resulted into very

high WER (reaching over 2000 in some cases). We removed these outliers by setting a threshold of 20 for WER for all dialects.

To ensure statistical rigor, we evaluated system differences at the utterance level, reporting 95% bootstrap confidence intervals (CIs), conducting paired tests to avoid sample mismatch, and accompanying all p-values with effect sizes (Cliff's δ). We further applied Holm–Bonferroni correction across the set of dialect–metric comparisons to control for family-wise error.

The results confirmed the Phase 1 findings: GPT correction reduced WER consistently across all dialects except Modern Standard Arabic (MSA), while improvements in SemDist were smaller but statistically significant for Najdi and Khaliji.

- Egyptian dialect showed a substantial reduction in WER (mean difference = 0.183, CI [0.050, 0.400]), with a small effect size (δ = 0.25). The SemDist improvement was negligible and not statistically significant, indicating GPT mainly reduced literal transcription errors without altering semantic fidelity.
- Hijazi dialect demonstrated modest WER gains (mean difference = 0.084, CI [−0.026, 0.199]) that did not reach significance after Holm adjustment. SemDist showed negligible effects (δ = 0.05), consistent with Whisper-Large already producing semantically faithful transcripts.
- Najdi dialect benefited from consistent improvements in both metrics: WER decreased by 0.055 (CI [0.018, 0.127]) and SemDist by 0.022 (CI [0.010, 0.033]). Both differences were statistically significant with negligible-to-small effect sizes, suggesting GPT correction systematically cleaned up errors without altering meaning.
- Khaliji dialect also showed clear gains: WER decreased by 0.078 (CI [−0.016, 0.176]) with a small effect size (δ = 0.16), while SemDist improved by 0.020 (CI [0.007, 0.033]), significant after correction. This indicates GPT post-processing improved both literal accuracy and semantic similarity for this dialect.
- MSA diverged from the pattern: GPT slightly increased WER (−0.084, CI [−0.283, 0.092]) and produced negligible SemDist change. Effect sizes were small and non-significant, consistent with the intuition that GPT "correction" may introduce unnecessary modifications when the baseline system already performs well on standardized language.

Overall, these analyses strengthen the Phase 1 conclusion: GPT correction yields consistent WER reductions in most dialects, with statistically validated improvements in Najdi and Khaliji. However, SemDist effects remain small, suggesting that Whisper-Large already preserves meaning relatively well, and GPT mainly corrects surface-level transcription errors. Importantly, the paired tests and

effect size analysis show that gains are robust but modest, and their practical utility may depend on application context. For MSA, where baseline quality is already high, GPT may risk overcorrection—an observation consistent with Naderi et al. [33], who reported occasional semantic drift when applying generative LLM correction to already high-quality transcriptions.

To assess the robustness of the observed improvements, we stratified the evaluation by utterance duration and signal-to-noise ratio (SNR) bins. Figure 5 (duration) and figure 6 (SNR) report mean WER with 95% bootstrap confidence intervals for Whisper-Large (Raw) and Whisper-Large+GPT across all dialects.

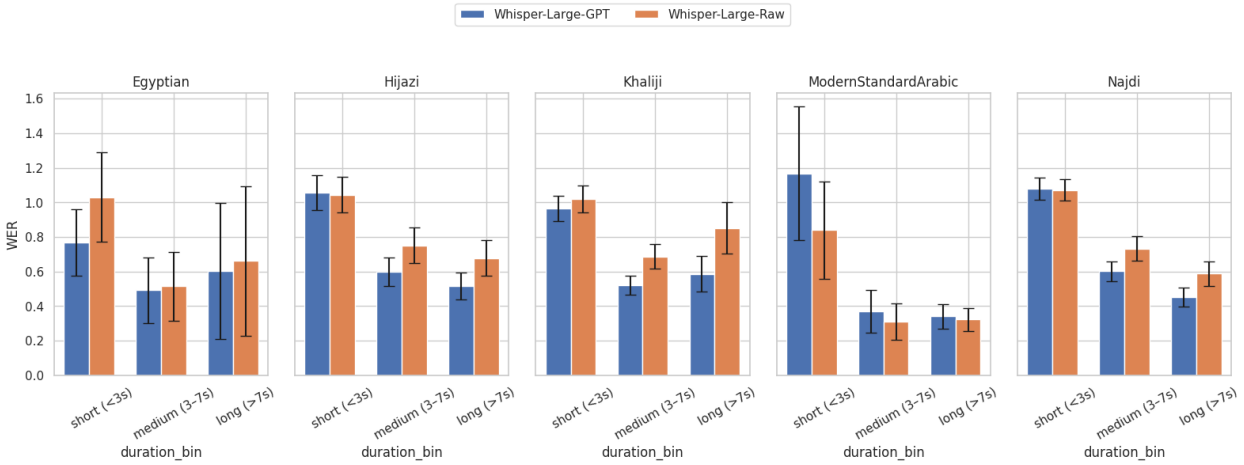


Fig. 5: WER by Duration Bin, Dialect, and System

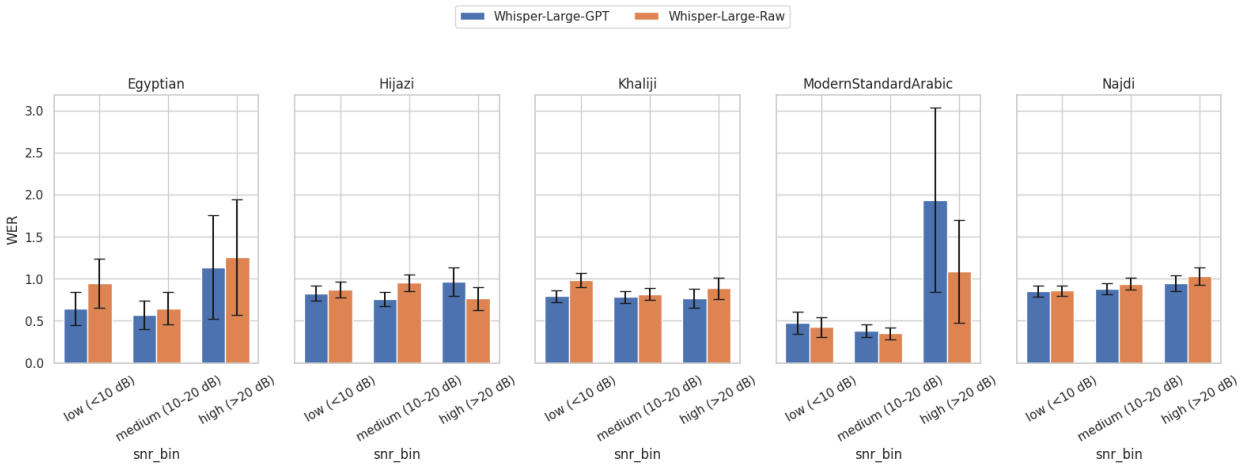


Fig. 6: WER by SNR Bin, Dialect, and System

When stratifying the results by utterance duration, we found that GPT post-processing did not provide uniform improvements across all bins. For short segments (<3s), gains were small and inconsistent: while Egyptian and Khaliji benefited modestly, MSA performance degraded, and Hijazi/Najdi showed negligible change. In contrast, for medium (3–7s) and long (>7s) segments, GPT consistently reduced

WER across dialects, with the largest relative improvements observed for Hijazi, Najdi, and Khaliji. This suggests that the LLM requires sufficient contextual information to apply meaningful corrections, and is less effective for short inputs where transcription errors may already be dominated by acoustic ambiguities rather than linguistic inconsistencies.

When stratifying by SNR, GPT also showed stable improvements in the low and medium ranges (<20 dB), confirming that it can mitigate some effects of noisy inputs. However, for high SNR (>20 dB), gains were smaller or inconsistent, with MSA showing degradation. This again indicates that LLM correction is most beneficial under challenging conditions, but can occasionally introduce unnecessary alterations when the base transcription is already relatively accurate.

The WER in all dialects apart from MSA can be considered relatively high, even after the possible gains from LLM post correction. Context-aware post-processing and few-shots LLM correction might increase the WER improvements. In addition, increasing the segments size could also provide more context to the LLM which may improve the results. Furthermore, as indicated in [34], dialect-pooled or fine-tuned models are needed to address resource scarcity.

5. Conclusion

This study presents a comprehensive evaluation of ASR performance for Arabic dialectal speech using Whisper and Wav2Vec2 models, supplemented by post-processing with large language models GPT and ALLAM. Our results demonstrate that Whisper, particularly in its larger variants, significantly outperforms Wav2Vec2 in zero-shot transcription across all tested dialects. However, both models show degradation in accuracy when transcribing non-standard or underrepresented dialects, such as Khaliji and Najdi, highlighting the limitations of current ASR systems in capturing the full linguistic diversity of Arabic.

Importantly, our findings confirm that large language models can meaningfully enhance ASR outputs. GPT, in particular, achieves reductions in both WER and Semantic Distance, indicating improvements not only in lexical correctness but also in semantic fidelity.

By combining quantitative and semantic evaluations, this study underscores the value of incorporating LLMs into ASR pipelines, especially for languages with rich dialectal variation. Future work could explore fine-tuning LLMs on dialectal corpora, integrating confidence-based editing mechanisms, and expanding evaluations to conversational or noisy speech scenarios. Ultimately, the integration of LLMs into ASR systems represents a promising direction for improving the accessibility and usability of Arabic speech technologies across dialects and domains.

Funding

The authors declare that this research received no external funding.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," OpenAI Technical Report, 2022.

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[3] Abdou and A. M. Moussa, "Arabic speech recognition: Challenges and state of the art," in *Computational Linguistics, Speech and Image Processing for Arabic Language*, 2019, pp. 1–27.

[4] A. Rahman, M. M. Kabir, M. F. Mridha, M. Alatiyyah, H. F. Alhasson, and S. S. Alharbi, "Arabic speech recognition: Advancement and challenges," *IEEE Access*, vol. 12, pp. 39689–39716, 2024.

[5] M. Afify, R. Sarikaya, H.-K. J. Kuo, L. Besacier, and Y. Gao, "On the use of morphological analysis for dialectal Arabic speech recognition," in *Proc. Interspeech 2006*, 2006, paper 1444-Mon2A2O.2, doi: 10.21437/Interspeech.2006-87.

[6] S. Alharbi *et al.*, "SADA: Saudi audio dataset for Arabic," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10286–10290.

[7] A. Ali *et al.*, "The MGB-2 challenge: Arabic multi-dialect broadcast media recognition," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 279–284, doi: 10.1109/SLT.2016.7846277.

[8] D. Wang, X. Wang, and S. Lv, "An overview of end-to-end automatic speech recognition," *Symmetry*, vol. 11, no. 8, p. 1018, 2019.

[9] N. Zerari, S. Abdelhamid, H. Bouzgou, and C. Raymond, "Bidirectional deep architecture for Arabic speech recognition," *Open Computer Science*, vol. 9, pp. 92–102, 2019.

[10] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Machine Learning (ICML)*, 2006, pp. 369–376.

[11] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.

[12] A. Waheed, B. Talafha, P. Sullivan, A. Elmadany, and M. Abdul-Mageed, "VoxArabica: A robust dialect-aware Arabic speech recognition system," in *Proc. ArabicNLP 2023 Workshop*, 2023, pp. 1–14.

[13] B. Talafha, A. Waheed, and M. Abdul-Mageed, "N-shot benchmarking of Whisper on diverse Arabic speech recognition," arXiv preprint, 2023.

[14] R. Sachdev, Z. Q. Wang, and C. H. H. Yang, "Evolutionary prompt design for LLM-based post-ASR error correction," arXiv preprint arXiv:2407.16370, 2024.

[15] R. Ma, M. Qian, M. Gales, and K. Knill, "ASR error correction using large language models," *IEEE Trans. Audio, Speech, Language Process.*, 2025.

[16] R. Ma, M. Qian, P. Manakul, M. Gales, and K. Knill, "Can generative large language models perform ASR error correction?" arXiv preprint arXiv:2307.04172, 2023.

[17] C. H. H. Yang, Y. Gu, Y. C. Liu, S. Ghosh, I. Bulyko, and A. Stolcke, "Generative speech recognition error correction with large language models and task-activating prompting," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8.

[18] Z. Min and J. Wang, "Exploring the integration of large language models into automatic speech recognition systems: An empirical study," in *Proc. Int. Conf. Neural Information Processing (ICONIP)*, Singapore, 2023, pp. 69–84.

[19] S. Li, C. Chen, C. Y. Kwok, C. Chu, E. S. Chng, and H. Kawai, "Investigating ASR error correction with large language model and multilingual 1-best hypotheses," in *Proc. Interspeech*, 2024, pp. 1315–1319.

[20] N. M. Matasyoh, R. A. Zeineldin, and F. Mathis-Ullrich, "Optimising speech recognition using LLMs: An application in the surgical domain," *Current Directions in Biomedical Engineering*, vol. 10, no. 1, pp. 45–48, Sep. 2024.

[21] S. Alharbi, R. Binmuqbil, A. Ali, R. Aloraini, S. Bari, A. Alowisheq, and Y. Alonaizan, "Leveraging LLM for augmenting textual data in code-switching ASR: Arabic as an example," in *Proc. SynData4GenAI*, 2024.

[22] T. Inoue, M. Elmahdy, A. Ali, and P. Bell, "Comparative study of end-to-end and hybrid ASR systems for Arabic broadcast speech," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 1085–1092.

[23] Y. Zhang, X. Wang, J. Xu, and H. Wang, "A comparative study of deep learning models for Arabic speech recognition," *IEEE Access*, vol. 10, pp. 10945–10953, 2022.

[24] S. Ding, A. Guo, A.-V. Nguyen, A. Vidon, J. Wang, and M. Zhou, "Data science capstone project: Robustness of Whisper and Wav2Vec2.0," 2023.

[25] Y. Wang, A. Alhmoud, and M. Alqurishi, "Open universal Arabic ASR leaderboard," arXiv preprint, 2024.

[26] A. C. Morris, V. Maier, and P. Green, "From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition," in *Proc. Interspeech*, 2004.

[27] A. Sahyoun and S. Shehata, "AraDiaWER: An explainable metric for dialectical Arabic ASR," in *Proc. 2nd Workshop on NLP Applications to Field Linguistics*, Dubrovnik, Croatia, 2023, pp. 64–73.

[28] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP/IJCNLP)*, 2019.

[29] M. Alian and A. Awajan, "Arabic semantic similarity approaches - review," in *Proc. Int. Arab Conf. Information Technology (ACIT)*, 2018, pp. 1–6, doi: 10.1109/ACIT.2018.8672665.

[30] T. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33, 2020, pp. 1877–1901.

[31] M. S. Bari *et al.*, "ALLaM: Large language models for Arabic and English," arXiv preprint, 2024.

[32] K. Almeman and M. Lee, "A comparison of Arabic speech recognition for multi-dialect vs. specific dialects," in *Proc. 7th Int. Conf. Speech Technology and Human-Computer Dialogue (SpeD)*, Cluj-Napoca, Romania, 2013.

[33] M. Naderi *et al.*, "Towards interfacing large language models with ASR systems using confidence measures and prompting," arXiv preprint arXiv:2407.19549, 2024.

[34] Ö. T. Özyilmaz, M. Coler, and M. Valdenegro-Toro, "Overcoming data scarcity in multi-dialectal Arabic ASR via Whisper fine-tuning," arXiv preprint arXiv:2506.02627, 2025.