When Does OmniASR Fail? A Fine-Grained Human Evaluation on Saudi Arabic Dialects

Hend Al-Khalifa

College of Computer and Information Sciences, King Saud University
Riyadh, Saudi Arabia
hendk@ksu.edu.sa

Abstract

Automatic Speech Recognition (ASR) evaluation has traditionally relied on Word Error Rate (WER), a metric that treats all errors equally and obscures critical failure modes. In this paper, we present a fine-grained human evaluation of Meta's recently released OmniASR system on Saudi Arabic dialects using the SADA dataset. Three trained annotators evaluated 103 audio samples, producing 264 annotations across two dimensions (comprehensibility and naturalness) while categorizing errors using a novel 10-category Arabic-specific error taxonomy. OmniASR achieved a mean WER of 42.2% and mean comprehensibility of 3.62/5, but exhibited a polarized performance pattern: 32.6% of transcriptions achieved perfect scores while 21.2% were essentially unusable. Error analysis reveals that hallucinations and deletions have the greatest negative impact on comprehensibility (−1.64 and −1.57 points respectively), roughly 6× more damaging than named entity errors. WER correlates with human comprehensibility ratings at a level comparable to inter-annotator agreement ($r = -0.679$ vs. pairwise annotator $r = 0.61$–0.67), but lacks the diagnostic granularity to reveal which error types drive quality degradation. These findings motivate the use of error-type-aware evaluation frameworks that complement WER with fine-grained analysis for Arabic ASR systems.

Keywords: Arabic ASR, OmniASR, Human Evaluation, Error Taxonomy, Saudi Dialects, SADA Dataset, Diglossia, WER

1. Introduction

Word Error Rate (WER) has been the dominant metric for evaluating Automatic Speech Recognition systems for decades. However, WER has significant limitations: it treats all errors equally, fails to capture semantic impact, and can obscure systematic failure patterns. These limitations are particularly pronounced for morphologically rich, dialectally diverse languages like Arabic, where a single morphological error may completely change the meaning of an utterance, while other errors may be cosmetic.

In November 2025, Meta released OmniASR (Keren et al., 2025), a multilingual ASR system supporting over 1,600 languages through a 7B-parameter Wav2vec 2.0 encoder combined with an LLM-based decoder. The system was designed for extensibility, enabling communities to add new languages with minimal data. While this represents a significant advancement in language coverage, the quality of transcription for individual languages, particularly Arabic dialects, remains understudied. The OmniASR paper reports aggregate Character Error Rates (CER) but provides limited analysis of error types or their perceptual impact on users.

Arabic presents unique challenges for ASR that go beyond those faced by well-resourced languages like English. The language exhibits diglossia (the coexistence of Modern Standard Arabic and regional spoken dialects), rich morphology with complex clitic attachment, and culturally sensitive content (e.g., religious phrases) that requires accurate transcription. These challenges motivate the need for evaluation frameworks that go beyond aggregate metrics.

This paper presents the first systematic human evaluation of OmniASR on Saudi Arabic dialects. We make three contributions: (1) a 10-category Arabic-specific error taxonomy that captures linguistically and culturally significant error types beyond what WER measures; (2) a human evaluation study with 264 annotations from 3 trained annotators on 103 audio samples from the SADA dataset (Alharbi et al., 2024); and (3) empirical findings demonstrating that error types have dramatically different impacts on perceived transcription quality, and that while WER correlates with human judgments at a level comparable to inter-annotator agreement, it cannot reveal which error types drive quality degradation.

The rest of the paper is organized as follows. Section 2 provides background on Arabic language characteristics relevant to ASR, including diglossia, morphological complexity, and code-switching. Section 3 reviews related work on Arabic ASR datasets, multilingual models, and evaluation methodologies. Section 4 describes our methodology, including the data sample, error taxonomy, annotation protocol, and ethical considerations. Section 5 presents our results on overall performance, error distribution, error impact analysis, inter-annotator agreement, and WER/CER analysis. Section 6 discusses the implications of our findings, and Section 7 concludes the paper.

Proceedings of SPEAKABLE @ LREC 2026, pages 118–124

2. Background

2.1 Arabic Diglossia and Dialectal Variation

Arabic is spoken by over 400 million people across more than 20 countries, making it the fifth most spoken language globally (Belinkov et al., 2019). However, 'Arabic' is not a single homogeneous language but rather a collection of varieties that exist in a state of diglossia (Ferguson, 1959). Native speakers acquire a regional spoken dialect as their first language, while Modern Standard Arabic (MSA), the standardized variety used in formal writing, news broadcasts, and official communication, is learned later through education (Badawi, 1973; Holes, 2004).

The differences between MSA and spoken dialects extend across all linguistic levels: phonology, morphology, syntax, and lexicon (Saiegh-Haddad, 2018). For example, the MSA word 'أريد' (ʔurīd, 'I want') becomes 'عايز' (ʕāyiz) in Egyptian, 'بدي' (biddi) in Levantine, and 'أبغى' (abġa) in Gulf Arabic. These are not merely pronunciation differences but entirely distinct lexical items. Such variation poses fundamental challenges for ASR systems, which must decide whether to transcribe what was actually said (dialect) or normalize to MSA, and the 'correct' choice depends on the use case.

Saudi Arabia alone encompasses multiple dialect groups: Najdi (central), Hijazi (western), and Gulf/Khaleeji (eastern), each with distinct phonological and lexical features. The SADA dataset we use in this study captures this diversity through content from Saudi television programs.

2.2 Arabic Morphological Complexity

Arabic morphology is based on a root-and-pattern system where most words derive from a triconsonantal root (e.g., k-t-b for writing-related words) combined with vowel patterns and affixes to create meaning (Boudelaa & Marslen-Wilson, 2013). This creates a high degree of morphological productivity but also ambiguity. Additionally, Arabic exhibits extensive use of clitics (attached pronouns, conjunctions, and prepositions that combine with base words). For example, 'وكتابي' (wa-kitāb-ī) encodes 'and my book' in a single orthographic unit.

For ASR, this morphological richness creates challenges at multiple levels. Segmentation errors (incorrect word boundaries) can create or destroy meaning. Clitic attachment errors change grammatical relationships. Dialectal morphological variants (e.g., different verb conjugation patterns) may be transcribed as MSA equivalents, altering the register and authenticity of the output. These considerations motivated several categories in our error taxonomy.

2.3 Arabic Code-Switching and Religious Language

Modern Arabic speech frequently involves code-switching with English, particularly in technical, business, and youth-oriented contexts. Terms like 'update,' 'email,' and 'meeting' are commonly inserted into Arabic discourse with varying degrees of phonological adaptation. ASR systems must decide whether to preserve these as English or attempt Arabic transliteration; both choices can create errors.

Additionally, religious phrases permeate everyday Arabic speech: greetings ('السلام عليكم', as-salāmu ʕalaykum), expressions of intent ('إن شاء الله', in shāʔ Allāh, 'God willing'), and responses to news ('الحمد لله', al-ḥamdu li-llāh, 'praise be to God'). These phrases carry cultural and religious significance, and errors in their transcription, particularly truncation or misspelling, can be perceived as disrespectful. This motivated our inclusion of 'Religious Phrases' as a distinct error category.

3. Related Work

Arabic ASR has benefited from several large-scale datasets in recent years. The MGB Challenge series introduced broadcast speech corpora covering multiple dialects (Ali et al., 2016), while the SADA dataset provides 668 hours of transcribed Saudi television content across Najdi, Hijazi, and Gulf dialects, with rich metadata on speaker characteristics (Alharbi et al., 2024). Benchmarking efforts such as the Open Universal Arabic ASR Leaderboard (Wang et al., 2024) have evaluated models including Whisper, MMS, and Wav2vec 2.0 variants, consistently finding that performance varies significantly across dialects, with MSA typically achieving lower error rates than dialectal varieties.

The most recent advance in multilingual ASR is OmniASR (Keren et al., 2025), which extends coverage to over 1,600 languages through scaled self-supervised pretraining with a 7B-parameter Wav2vec 2.0 encoder combined with an LLM-inspired decoder. The system achieves character error rates below 10% for 78% of supported languages and introduces zero-shot capabilities for adding new languages with minimal paired examples. However, its evaluation focuses primarily on aggregate CER metrics across languages rather than detailed error analysis for specific language families like Arabic.

This reliance on aggregate metrics reflects a broader limitation in ASR evaluation. Researchers have long recognized that WER treats all errors equally regardless of their semantic impact. Alternative approaches include Semantic Error Rate, which attempts to weight errors by importance (McCowan et al., 2004), and human evaluation frameworks from machine translation

that separate adequacy from fluency (Koehn & Monz, 2006), an approach we adapt here as comprehensibility and naturalness. More recently, calls for 'slice-aware' metrics have emphasized the need to reveal performance variation across demographic groups and linguistic conditions (Tatman, 2017). Despite these advances, fine-grained error taxonomies for Arabic ASR remain underexplored, with most evaluations continuing to rely solely on WER or CER. Our work addresses this gap by introducing an Arabic-specific error taxonomy and correlating error types with human quality judgments.

4. Methodology

4.1 Data

We used the SADA dataset (Alharbi et al., 2024), which contains 668 hours of transcribed Arabic audio from Saudi television programming, covering Najdi, Hijazi, and Gulf dialects with metadata on speaker characteristics. From this dataset, we sampled 103 audio files for human evaluation, stratified by dialect to ensure representation across Arabic varieties. Our annotated sample included Najdi (57.3%), Gulf/Khaliji (11.7%), Hijazi (10.7%), Egyptian (6.8%), Levantine (2.9%), and other varieties including MSA, Sudanese, Tunisian, and Saidi (10.6%). Audio files ranged from short utterances (4 words) to longer segments (up to 19 words), with an average length of approximately 10 words per utterance.

4.2 ASR System

We used Meta's OmniASR system via the HuggingFace Spaces interface (facebook/omniasr-transcriptions). OmniASR employs a 7B-parameter Wav2vec 2.0 encoder pretrained on massive multilingual speech data, combined with an LLM-based decoder that leverages language model priors for improved transcription (Keren et al., 2025). We used the default inference settings without language conditioning, simulating a realistic zero-shot usage scenario where users may not specify the dialect.

4.3 Error Taxonomy

We developed a 10-category error taxonomy specifically designed for Arabic ASR evaluation, drawing on linguistic literature on Arabic diglossia, morphology, and sociolinguistic variation. The taxonomy captures both general ASR errors (deletions, insertions, substitutions) and Arabic-specific phenomena (dialect-MSA substitution, religious phrases, and morphological errors). Table 1 presents the complete taxonomy with examples.

Error Type	Description	Example	Transliteration
Hallucination	Words inserted not spoken	العرض → متاح العرض متاح مجانا	*'available'* → *'available for free'*
Deletion	Spoken words omitted	لا أوافق → أوافق	*'I do NOT agree'* → *'I agree'*
Substitution	Semantic word replacement	التذكرة → صالحة التذكرة صعبة	*'ticket is valid'* → *'ticket is difficult'*
Segmentation	Incorrect word boundaries	بالرغم → بال رغم	*'despite'* → *incorrectly split*
Morphology	Clitic/conjugation errors	كتابي → كتاب	*'my book'* → *'book'*
Named Entities	Names, places, orgs	الرياض → الرواد	*'Riyadh'* → *'al-Rawad'*
Religious Phrases	Islamic expressions	إن شاء الله → إنشاء الله	*'God willing'* → *misspelled*
Dialect↔MSA	Register substitution	عايز → أريد	*Egyptian dialect* → *MSA*
Code-switching	Foreign word errors	GitHub → جتهاب	*English* → *Transliterated Arabic*
Numbers	Digits, dates, prices	١٢ مارس → ٢١ مارس	*'March 12'* → *'March 21'*

Table 1: Arabic ASR Error Taxonomy. Arrow (→) indicates ground truth → erroneous output.

4.4 Annotation Protocol

Three senior undergraduate IT students who had completed an NLP course and are native Arabic speakers were trained as annotators. Training included detailed Arabic-language guidelines with the error taxonomy, examples, rating scale anchors, and practice items with feedback. For each audio-transcription pair, annotators: (1) listened to the audio without viewing the transcription; (2) reviewed the OmniASR transcription; (3) rated Comprehensibility (1-5): How easily can the transcription be understood? (4) rated Naturalness (1-5): How natural does the Arabic text sound? (5) selected all applicable error types from the taxonomy. Annotation was conducted using Google Forms, with 62 audio files evaluated by all three annotators to enable inter-annotator agreement analysis.

5. Results

5.1 Overall Performance

OmniASR achieved moderate overall performance on our sample: mean

comprehensibility of 3.62/5 (σ=1.35) and mean naturalness of 4.05/5 (σ=1.26). However, the distribution exhibits a polarized pattern with elevated frequency at both extremes (Figure 1). Perfect scores (5/5 on both dimensions) were achieved on 32.6% of transcriptions, while 21.2% scored ≤2 on comprehensibility, essentially unusable for any downstream application. Although the middle range (scores 2–4) accounts for 54% of ratings, the proportion of extreme scores is notably higher than expected under a normal distribution. This polarized pattern suggests OmniASR either succeeds well or fails catastrophically, rather than degrading gradually.

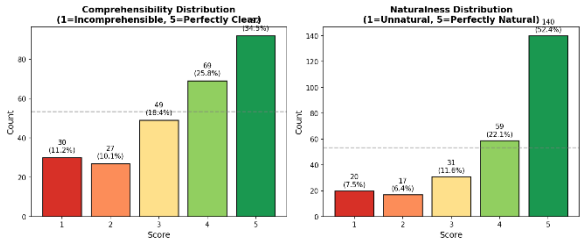

Figure 1: Distribution of comprehensibility and naturalness scores across 264 evaluations.

5.2 Error Type Distribution

Figure 2 shows the distribution of error types across 442 total error instances. The most common errors were semantic substitutions (22.9%), deletions (22.4%), hallucinations (18.3%), and segmentation errors (16.5%). Morphological errors accounted for 11.1%. Arabic-specific error types were relatively rare: named entities (5.4%), religious phrases (1.4%), dialect↔MSA substitution (0.7%), numbers (0.7%), and code-switching (0.5%).

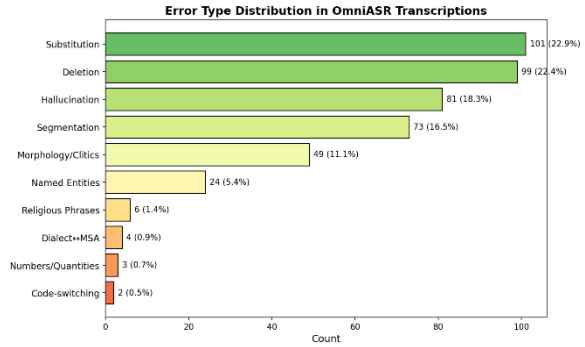

Figure 2: Distribution of error types in OmniASR transcriptions (n=442 error instances).

5.3 Error Impact Analysis

Critically, not all errors affect comprehensibility equally (Figure 3). To quantify impact, we computed the difference in mean comprehensibility between transcriptions where each error type was present versus absent. Because our data include repeated judgments by individual raters on multiple audio files, these mean-difference estimates should be interpreted as descriptive rather than inferential; a mixed-effects regression with random intercepts for annotator and audio file would be needed for formal hypothesis testing. With this caveat, hallucinations show the greatest negative impact (−1.64 points when present vs. absent), followed by deletions (−1.57), segmentation errors (−1.33), and substitutions (−1.18). Notably, named entity errors have minimal impact on human comprehensibility ratings (−0.27), suggesting readers can often infer correct names from context. However, this low impact on comprehensibility should not be taken to mean that named entity errors are inconsequential: for downstream NLP tasks such as named entity recognition, information extraction, or question answering, entity errors may be among the most damaging, effectively inverting the impact ranking observed here.

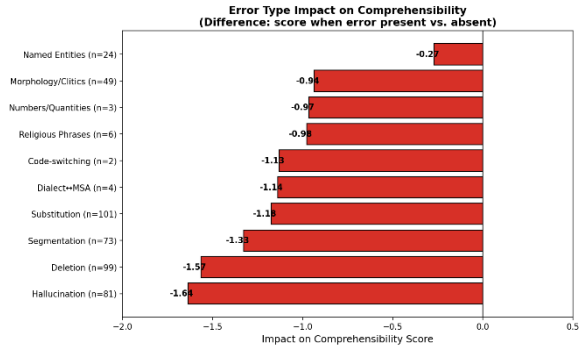

Figure 3: Impact of each error type on comprehensibility.

5.4 Error Count and Co-occurrence

We found a strong negative correlation between error count and comprehensibility (Pearson r = −0.776, p < 0.001; Figure 4). Transcriptions with zero errors averaged 4.86/5, declining to 4.20/5 for one error, 3.26/5 for two errors, 2.53/5 for three errors, and 1.84/5 for four errors. The most common error pairs were deletion+hallucination (58 co-occurrences), deletion+substitution (46), and hallucination+segmentation (45).

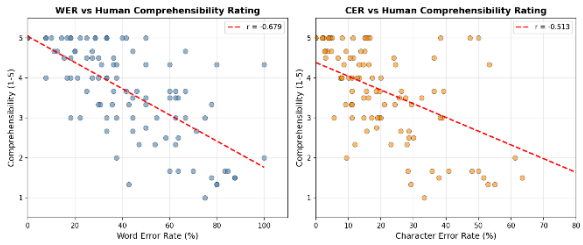

Figure 4: Mean comprehensibility by number of errors per transcription.

5.5 Dialect Effects

Performance varied by dialect. Based on human ratings, Najdi dialect achieved the highest comprehensibility (3.83), followed by Gulf Arabic (3.44) and Egyptian (3.40). MSA samples scored perfectly (5.00) but the sample size was too small (n=4) for reliable conclusions.

5.6 Inter-Annotator Agreement

For the 62 files evaluated by all three annotators, we computed pairwise Pearson correlations. For comprehensibility, correlations ranged from 0.61 to 0.67 (moderate-to-good agreement). Within ±1 point agreement was achieved on 79.6% of pairwise comparisons, with exact agreement on 40.9%. Naturalness showed more variation (r = 0.26 to 0.59), consistent with its more subjective nature. These agreement levels support the reliability of the annotation task while acknowledging inherent subjectivity in quality judgments.

5.7 Automatic Metric Analysis (WER/CER)

To complement our human evaluation, we computed Word Error Rate (WER) and Character Error Rate (CER) by aligning OmniASR transcriptions with SADA ground truth for 100 matched audio files. OmniASR achieved a mean WER of 42.2% (σ=24.0%, median=37.5%) and mean CER of 21.1% (σ=17.3%, median=16.1%). The WER distribution reveals substantial variation: 19% of files achieved excellent performance (WER <20%), while 37% showed poor performance (WER >50%).

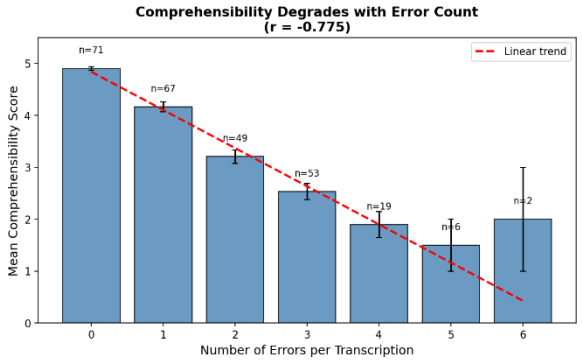

Figure 5: Correlation between automatic metrics (WER/CER) and human comprehensibility ratings.

We examined the correlation between automatic metrics and human judgments (Figure 5). WER showed a strong negative correlation with comprehensibility (r = −0.679), while CER showed a moderate correlation (r = −0.513). Notably, the WER–comprehensibility correlation (|r| = 0.679) is comparable to inter-annotator agreement for comprehensibility (pairwise r = 0.61–0.67), indicating that WER predicts human judgments approximately as well as individual annotators predict each other. This suggests that WER is a reasonable single-number proxy for overall perceived quality. However, WER by design treats all word-level errors as equivalent and cannot reveal which error types are most damaging. The value of our fine-grained taxonomy lies precisely in this diagnostic capability: identifying that hallucinations and deletions are 6× more impactful than named entity errors, information that no aggregate metric can provide regardless of its correlation with human judgments.

6. Discussion

6.1 The Polarized Performance Problem

Perhaps our most noticeable finding is OmniASR's polarized performance distribution: approximately one-third of transcriptions are essentially perfect, while one-fifth are essentially unusable. While the majority of scores (54%) fall in the middle range (2–4), the elevated frequency at both extremes is notable. This pattern would be obscured by aggregate metrics like mean WER (42.2%) or CER (21.1%). For practical applications, users need to know not just average quality but the probability of catastrophic failure.

This polarization likely reflects how neural ASR models process speech. When the acoustic signal clearly matches patterns in training data, the model produces confident, accurate transcriptions. When there is a mismatch due to unfamiliar dialect features, acoustic conditions, or speaker characteristics, the model may 'fall off a cliff' rather than degrading gracefully. The strong correlation between error count and comprehensibility (r = −0.776) supports this interpretation: errors cluster together in problematic utterances rather than being distributed uniformly.

6.2 Why Hallucinations and Deletions Are Most Damaging

Our analysis revealed that hallucinations (−1.64 impact) and deletions (−1.57) are by far the most damaging error types for comprehensibility, roughly 6× more impactful than named entity errors (−0.27). Hallucinations are particularly problematic because they introduce false information that readers have no way to detect or correct. In high-stakes applications (medical, legal, journalistic), hallucinated content could have serious consequences. The prevalence of hallucinations (18.3% of errors) suggests that OmniASR's LLM-based decoder may be over-generating content to create fluent output.

Deletions are damaging because they remove information entirely, potentially changing meaning. The Arabic sentence 'لا أوافق' (I do not agree) becomes 'أوافق' (I agree) if the negation particle is deleted. This type of semantically critical deletion, where function words such as negation markers are omitted, was observed in our data and illustrates how deletions can invert meaning entirely. The high co-occurrence of deletions with hallucinations (58 pairs) suggests a common failure mode: the model skips content it cannot confidently transcribe and compensates by generating plausible filler.

6.3 The Fluency Trap

OmniASR achieved notably higher naturalness scores (mean 4.05) than comprehensibility scores (mean 3.62). This gap reveals that the LLM-based decoder produces fluent, natural-sounding Arabic text that may nonetheless be semantically incorrect. This 'fluency trap' has been documented in neural machine translation (Martindale & Carpuat, 2018). For ASR evaluation, this finding argues strongly for separating fluency/naturalness from accuracy/comprehensibility in evaluation frameworks. An important question for future work is whether this fluency trap is specific to LLM-based decoding. CTC-based models, which decode frame-by-frame without autoregressive language model priors, would likely produce less fluent but potentially more faithful transcriptions. Comparing LLM-based and CTC-based architectures using our evaluation framework could reveal whether there is a systematic trade-off between naturalness and comprehensibility across decoding strategies.

6.4 WER and the Role of Fine-Grained Analysis

Our WER/CER analysis reveals a nuanced picture of WER's role in ASR evaluation. The WER–comprehensibility correlation ($r = -0.679$) is comparable to inter-annotator agreement on comprehensibility (pairwise $r = 0.61–0.67$), indicating that WER predicts human judgments at a level on par with the agreement among individual annotators. WER therefore serves as a reliable single-number summary of overall quality. However, the value of our error taxonomy is not in replacing WER but in complementing it with diagnostic information that WER by design cannot provide. WER treats all word-level errors as equivalent, yet our analysis shows that hallucinations and deletions are roughly 6× more damaging to comprehensibility than entity errors. A weighted WER scheme informed by these impact differences could provide a more informative automatic metric, and our taxonomy offers the empirical basis for developing such weightings.

6.5 Implications for Arabic ASR Evaluation

Our findings have several implications for Arabic ASR evaluation: (1) Weighted Error Metrics: evaluation metrics should weight errors by their perceptual impact; (2) Distributional Reporting: beyond mean WER, report the proportion of catastrophic failures and successes; (3) Dialect-Specific Evaluation: performance should be reported separately for MSA and each major dialect group; (4) Arabic-Specific Error Categories: track morphological errors, religious phrase accuracy, and dialect-MSA substitution; (5) Task-Dependent Impact Profiles: the error impact ranking we report reflects human comprehensibility, but downstream applications may exhibit entirely different sensitivity profiles. For instance, named entity errors had minimal impact on comprehensibility (−0.27) yet could be highly damaging for entity-centric tasks such as information extraction and question answering. Future work should evaluate the same transcriptions through downstream NLP pipelines to produce task-specific impact rankings that complement our human-centered analysis.

7. Conclusion

We presented the first fine-grained human evaluation of Meta's OmniASR system on Saudi Arabic dialects, introducing a 10-category error taxonomy designed for Arabic's linguistic and cultural characteristics. OmniASR achieved 42.2% WER and 3.62/5 mean comprehensibility, but exhibited a polarized performance pattern with 32.6% perfect transcriptions and 21.2% unusable ones. Error analysis reveals that hallucinations and deletions are approximately 6× more damaging than named entity errors for human comprehensibility, though the impact ranking may differ for downstream NLP tasks. While WER correlates with human judgments at a level comparable to inter-annotator agreement ($r = -0.679$ vs. pairwise annotator $r = 0.61–0.67$), it cannot identify which error types drive quality degradation. Our error taxonomy complements WER by providing this diagnostic granularity, supporting the development of error-type-aware evaluation frameworks for Arabic ASR.

8. Limitations and Ethical Considerations

Our study has several limitations. First, our sample size (103 audio files, 264 evaluations) is modest; a larger-scale study would provide more robust estimates, particularly for rare error types and dialect-specific breakdowns. Second, our annotators were senior undergraduate students who had completed an NLP course and received task-specific training rather than professional linguists; inter-annotator agreement for comprehensibility ($r = 0.61–0.67$) falls within ranges reported in similar ASR evaluation studies, but expert annotators might identify errors more consistently. Third, our error impact estimates are based on mean differences between present/absent conditions; a mixed-effects regression with random intercepts for annotator and audio file would provide more rigorous statistical inference, and we note this as a direction for follow-up analysis. Fourth, we evaluated only OmniASR, an LLM-based system; comparative evaluation against CTC-based models (e.g., Wav2Vec 2.0 CTC) and attention-based models (e.g., Whisper) would reveal whether the fluency trap and error type distributions we observe are architecture-specific. Finally, our sample was primarily Saudi dialects

with some Egyptian content; evaluation on Levantine and Maghrebi dialects would further test the generalizability of our error taxonomy.

The annotation task was conducted as part of a graded course assignment in an NLP course. Grades were based on task completion and annotation quality (consistency and thoroughness) rather than on achieving particular results. The SADA dataset used in this study is publicly available under CC BY-NC-SA 4.0 license.

9. References

Ali, A., Bell, P., Glass, J., Messaoui, Y., Mubarak, H., Renals, S., & Zhang, Y. (2016). The MGB-2 Challenge: Arabic Multi-Dialect Broadcast Media Recognition. In Proceedings of the 2016 IEEE Spoken Language Technology Workshop (SLT) , pp. 279-284. IEEE. https://doi.org/10.1109/SLT.2016.7846277

Alharbi, S., Alowisheq, A., Tüske, Z., Darwish, K., Alrajeh, A., Alrowithi, A., Bin Tamran, A., Ibrahim, A., Aloraini, R., Alnajim, R., Alkahtani, R., Almuasaad, R., Alrasheed, S., Alsubaie, S., & Alonaizan, Y. (2024). SADA: Saudi Audio Dataset for Arabic. In Proceedings of ICASSP 2024 - IEEE International Conference on Acoustics, Speech and Signal Processing , pp. 10286-10290. IEEE. https://doi.org/10.1109/ICASSP48485.2024.10446243

Badawi, E. S. (1973). Mustawayat al-'arabiyya al-mu'asira fi Misr [Levels of Contemporary Arabic in Egypt]. Cairo: Dar al-Ma'arif.

Belinkov, Y., Barrón-Cedeño, A., Magidow, A., Shmidman, A., & Romanov, M. (2019). Studying the history of the Arabic language: Language technology and a large-scale historical corpus. Language Resources and Evaluation , 53(4), 771-805. https://doi.org/10.1007/s10579-019-09460-w

Boudelaa, S., & Marslen-Wilson, W. D. (2013). Morphological structure in the Arabic mental lexicon: Parallels between standard and dialectal Arabic. Language and Cognitive Processes , 28(10), 1453-1473. https://doi.org/10.1080/01690965.2012.719629

Ferguson, C. A. (1959). Diglossia. Word , 15(2), 325-340. https://doi.org/10.1080/00437956.1959.11659702

Holes, C. (2004). Modern Arabic: Structures, Functions, and Varieties (Revised ed.). Georgetown University Press.

Keren, G., Kozhevnikov, A., Meng, Y., Ropers, C., Setzler, M., Wang, S., Adebara, I., Auli, M., Balioglu, C., Chan, K., Cheng, C., Chuang, J., Droof, C., Duppenthaler, M., Duquenne, P.-A., Erben, A., Gao, C., Gonzalez, G. M., Lyu, K., ... Yates, S. (2025). Omnilingual ASR: Open-Source Multilingual Speech Recognition for 1600+ Languages. arXiv preprint arXiv:2511.09690 . https://arxiv.org/abs/2511.09690

Koehn, P., & Monz, C. (2006). Manual and Automatic Evaluation of Machine Translation between European Languages. In Proceedings of the Workshop on Statistical Machine Translation , pp. 102-121. Association for Computational Linguistics.

Martindale, M. J., & Carpuat, M. (2018). Fluency Over Adequacy: A Pilot Study in Measuring User Trust in Imperfect MT. In Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (AMTA) , Vol. 1, pp. 13-25.

McCowan, I., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P., & Bourlard, H. (2004). On the Use of Information Retrieval Measures for Speech Recognition Evaluation. IDIAP Research Report 04-73 . IDIAP.

Saiegh-Haddad, E. (2018). MAWRID: A Model of Arabic Word Reading in Development. Journal of Learning Disabilities , 51(5), 454-462. https://doi.org/10.1177/0022219417720460

Tatman, R. (2017). Gender and Dialect Bias in YouTube's Automatic Captions. In Proceedings of the First ACL Workshop on Ethics in Natural Language Processing , pp. 53-59. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1606

Wang, Y., Alhmoud, A., & Alqurishi, M. (2024). Open Universal Arabic ASR Leaderboard. https://arxiv.org/abs/2412.13788