Beyond Conventional ASR Metrics: Evaluating Semantic Preservation in Low-Resource Speech Recognition

Nurul Emilia Diyana Abdul Malik[1], Mohd Izzuddin Mohd Tamrin[2],
Airul Azizan Zainudin[3], Muhamad Ariff Ibrahim[4], Mohd Zulfaezal Che Azemin[5]*
Kulliyyah of Allied Health Sciences, International Islamic University Malaysia, Kuantan, Malaysia[1, 4, 5]
Department of Information Systems-Kulliyyah of Information and Communication Technology,
International Islamic University Malaysia, Gombak, Malaysia[2]
Department of Professional Nursing Studies, Kulliyyah of Nursing, International Islamic University Malaysia[3]

***Abstract*—Word error rate (WER) and character error rate (CER) quantify lexical discrepancies in automatic speech recognition (ASR) but do not directly indicate whether the intended message remains intact. This exploratory study compared lexical accuracy with human-rated preservation of essential meaning in Malay-English code-switched conversational speech. Four complete inference pipelines were evaluated on 30 purposively selected windows from one recording (826 reference words): Mesolitica Malaysian Whisper Large-v3-Turbo-v3 (Pipeline 1), OpenAI Whisper Large-v3-Turbo through faster-whisper (Pipeline 2), Qwen3-ASR-1.7B with a combined Malay-English contextual prompt (Pipeline 3), and the same Qwen3-ASR model without that prompt (Pipeline 4). Three evaluators independently assessed 120 reference-hypothesis pairs; a two-of-three majority constituted the semantic endpoint. Pipeline 4 had the lowest WER (29.18%) and CER (20.60%), whereas Pipeline 1 had the highest observed majority-rated semantic-preservation proportion (53.3%, versus 36.7% for Pipeline 4). The overall pipeline effect was significant for WER (Friedman chi-square (3) = 15.64, $p = 0.0013$) and CER (chi-square (3) = 18.17, $p < 0.001$), but not for semantic preservation (Cochran's $Q = 5.57$, $p = 0.134$); no paired semantic contrast remained significant after Holm adjustment. The Pipeline 1–Pipeline 4 semantic difference was 16.7 percentage points (paired-window bootstrap 95% confidence interval: −6.7 to 40.0 percentage points). Overall Fleiss' kappa was 0.529. Different descriptive rankings were therefore observed, but this single-recording sample does not establish significant differences in semantic performance or general model superiority. These findings support further evaluation of human semantic judgments alongside lexical metrics, with independent validation across recordings, speakers, and inference implementations.**

Keywords*—Automatic speech recognition; Malay-English code-switching; low-resource speech; word error rate; character error rate; semantic preservation; human evaluation; exploratory benchmarking

I. Introduction

Automatic speech recognition (ASR) has advanced through multilingual pretraining, self-supervised representation learning, and weak supervision. These approaches are relevant to languages and speech domains with limited labelled data, where supervised ASR development can be constrained by the availability of reliable transcriptions [1], [2]. Cross-lingual models such as XLS-R and benchmarks such as FLEURS have demonstrated the utility of multilingual speech representations [3], [4]. Whisper extended this paradigm through large-scale weak supervision [5], while more recent foundation-model-based systems, including Qwen3-ASR, have expanded multilingual transcription capabilities [6].

Despite these advances, low-resource and code-switched speech remains difficult. Recent work on low-resource Malay speech, including automated assessment of children's speech fluency, illustrates the continuing need for adapted multilingual ASR approaches when labelled speech resources are limited [7]. Code-switching introduces rapid language alteration, informal grammar, borrowed lexical forms, pronunciation variability, and strong dependence on local sociolinguistic context [8], [9]. Work on multilingual code-switched ASR in data-sparse settings has shown that performance is highly sensitive to the availability and type of in-domain speech and text [10]. Recent Southeast Asian studies have specifically addressed Malay-English code-switching, including synthetic code-switch adaptation and asynchronous text-speech adaptation for Whisper-based systems [11], [12]. More generally, recent foundation models can remain challenged by intra-sentential code-switching even when they possess strong multilingual pretraining [13].

ASR performance is nevertheless still reported predominantly using word error rate (WER), character error rate (CER), or closely related edit-distance measures. These measures are indispensable for quantifying literal deviations from a reference transcript, but they assign errors according to token-level edits rather than the importance of those errors to communicated meaning. This limitation has long been recognized: alternative measures such as word information loss were proposed because sequence restoration and information transmission are not identical objectives [14], and optimizing WER has been shown to be misaligned with downstream speech translation quality in some settings [15].

More recent ASR evaluation research has attempted to capture meaning directly. Semantic Distance was proposed as an embedding-based complement to WER and was reported to align more closely with downstream spoken-language understanding [16]. Follow-up work showed stronger

*Corresponding author.

correspondence between semantic distance and user judgments of transcription quality than WER alone [17]. Aligned Semantic Distance similarly demonstrated that transcripts with comparable WER can differ materially in semantic quality [18] and that semantic measures may better track human perception and downstream task performance [19]. SeMaScore further combined recognition error and similarity information in an evaluation measure that corresponded well with expert assessment [20]. Recent Whisper-specific analysis has also shown that aggregate WER can mask systematic behaviour across language-resource levels [21].

This literature raises an empirical question: when ASR systems differ in lexical accuracy, do their outputs receive the same ordering when evaluated for preservation of essential meaning? We examined this question in an exploratory within-recording case study of Malay-English code-switched speech. Four complete ASR inference pipelines processed the same 30 acoustic windows and were compared using WER, CER, and three independent human semantic judgments per output. The contribution is a transparent matched-item evaluation of potentially differing descriptive rankings rather than a new ASR architecture, a causal assessment of regional model adaptation, or a generalizable benchmark of Malay-English speech recognition.

II. Related Work

A. Low-Resource and Multilingual ASR

Low-resource ASR is characterized not only by limited hours of labelled speech but also by mismatches in domain, speaker population, orthography, and available linguistic resources [1]. Large multilingual models mitigate some of these limitations by sharing representations across languages. XLS-R showed that scaling cross-lingual self-supervised pretraining across 128 languages can improve recognition in multilingual and low-resource conditions [3]. FLEURS provided a parallel 102-language benchmark designed to assess few-shot and universal speech representations across diverse resource levels [4]. Empirical work has nevertheless emphasized that large pretrained models still benefit from in-domain adaptation and that simply adding out-of-domain data does not guarantee improved low-resource performance [2].

Whisper is especially relevant because its multilingual weakly supervised pretraining provides broad zero-shot coverage [5]. In the Malay context, Zhang et al. used a pre-finetuned Mesolitica Malaysian Whisper-medium model as part of a low-resource children's speech-fluency pipeline [7]. That study provides published precedent for the Mesolitica Whisper family in Malay ASR, although it did not evaluate the exact Large-v3-Turbo-v3 checkpoint used in the present work.

B. Code-Switched and Malay-English ASR

Code-switched ASR combines the challenges of multilingual recognition with within-utterance language switching. A review of bilingual and multilingual code-switching systems identified data sparsity and large variation in recognition accuracy as persistent problems [8], while a recent systematic review of end-to-end code-switched ASR mapped continuing gaps in datasets, model design, and evaluation across language pairs [9]. In South African code-switched ASR, both in-domain and out-of-domain resources improved performance, but in-domain resources were generally more effective, underscoring the importance of local speech characteristics [10].

For Southeast Asian language pairs, Nguyen and Tran reported that synthetic phrase-level code-switching could improve ASR performance without a large corpus of real code-switched speech [11]. Their subsequent AsyncSwitch framework also reported improvements for Whisper-based Malay-English ASR, including a 9.02% relative WER reduction [12]. These studies provide regional context but emphasize lexical recognition performance. The present investigation added an independently rated binary semantic endpoint to examine whether lexical and human-judged outcomes yielded the same descriptive ordering on matched excerpts; it was not designed to establish the broader validity of a new semantic metric.

C. Beyond WER: Semantic Evaluation of ASR

WER remains valuable because it is interpretable, reproducible, and directly tied to substitutions, deletions, and insertions. However, a substitution of a function word and a substitution that reverses a key clinical, numerical, or factual concept each count as a single token error. Conversely, multiple lexical differences may leave the intended message intact. This mismatch between token error and informational consequence motivated early alternatives to WER [14], [15] and has become more salient as ASR is increasingly used as an upstream component for summarization, translation, retrieval, and language understanding.

Automated semantic metrics provide additional approaches to measuring meaning-related transcription differences. Prior studies have reported relationships between Semantic Distance and spoken-language understanding or user judgments [16], [17], while Aligned Semantic Distance and SeMaScore have been evaluated as alternative semantic-sensitive metrics [18]–[20]. These published findings motivated the inclusion of a human judgment endpoint in the present study. Semantic Distance, Aligned Semantic Distance, and SeMaScore were not calculated on the present 30 excerpts, and their agreement with our human judgments is not an outcome of this investigation.

III. Methodology

A. Study Design and Evaluation Corpus

A within-item, exploratory comparison was undertaken using a single Malay-English code-switched conversational recording. The source 'report' worksheet contained 289 candidate speech windows with reference transcripts and outputs from four pipelines. Thirty windows were purposively identified by non-empty Item labels to span recording positions and utterance durations; the other candidates were not included in the formal semantic evaluation. The same 30 windows (826 reference words) were used for both lexical and semantic analyses. Because the excerpts were not selected randomly or stratified across independent recordings, the results describe this selected within-recording sample rather than an unbiased estimate of performance for Malay-English speech generally.

The selected excerpts covered original window identifiers 42–320 and start times 598.5–4083.2 s; median duration was

13.6 s (mean 10.9 s; range 2.5–17.0 s). All reference excerpts contained scorable alphanumeric text, and all four pipelines returned non-empty hypotheses for the selected items. The reference texts were provided as 'ground truth' in the evaluation worksheet. However, the available materials do not establish the original transcription process, whether the reference text was checked against the recording by independent annotators, or whether any disagreements were adjudicated. Accordingly, these texts are termed 'reference transcripts' without assuming an independently verified gold standard.

B. Audio Preprocessing and Common Speech-Window Generation

Source audio was converted by FFmpeg to 16-kHz mono pulse-code-modulated waveform data. Common ASR windows were generated once and reused across pipelines, ensuring identical audio start and end points for each comparison. Silero voice activity detection used a probability threshold of 0.50, minimum speech duration of 150 ms, minimum silence duration of 500 ms, and 200 ms of speech padding. Nearby regions were combined to target windows of approximately 5–15 s; short intervals were merged where feasible, with a 2.5-s minimum duration and a maximum span of 20 s. These settings characterize the supplied V5 scripts rather than an independently archived execution environment.

Available Pyannote RTTM speaker-diarization information was mapped to each ASR window as metadata (dominant speaker, speakers present, and overlap fraction). Diarization boundaries did not determine audio segmentation. This separation was intended to avoid unstable sub-second ASR inputs while maintaining the same acoustic excerpts for every pipeline; it does not remove within-recording clustering by speaker, topic, or acoustic conditions.

C. ASR Pipelines

Four complete inference pipelines were compared on common speech windows (Table I). The input recording, audio normalization, temporal windows, and broadly similar Malay/English language-restriction and quality-control policies were shared. Model checkpoints, inference libraries, generation interfaces, precision pathways, and token limits were not identical. Consequently, this design assessed end-to-end pipeline outputs; it did not isolate the effect of Malay-specific model adaptation or any single backend component. The available V5 source scripts were audited for specified defaults, but exact executed model revision identifiers, installed package versions, driver versions, and original hardware manifests could not be verified from the submitted evaluation files.

Pipeline 1 used 'mesolitica/Malaysian-whisper-large-v3-turbo-v3' with Hugging Face WhisperProcessor/ WhisperForConditionalGeneration; the source script specified deterministic generation with one beam and a maximum of 256 new tokens. Pipeline 2 used 'large-v3-turbo' through faster-whisper/CTranslate2, with default float16 computation, beam size 1, best-of 1, temperature 0, and a maximum of 256 new tokens. Pipelines 3 and 4 used 'Qwen/Qwen3-ASR-1.7B-hf' through the Transformers multimodal model interface, with deterministic generation and a maximum of 512 new tokens. The Qwen loading script selected bfloat16 when supported and float16 otherwise.

Pipeline 3 used one combined contextual prompt: 'Malaysian conversation in Malay and English. Speech may contain Malay-English code-switching (Manglish). Preserve the spoken Malay and English words verbatim. Do not translate. Use Latin script unless a proper name requires otherwise.' Pipeline 4 used the same Qwen model with that prompt disabled. This was a combined-prompt versus no-prompt comparison; it was not a component-wise ablation of language identification, code-switch context, verbatim instruction, or script guidance. Under the restricted-auto policy, the automatic output was retained when Malay/English language identification and relevant script checks were acceptable; otherwise, Malay- and English-constrained hypotheses were evaluated using a quality-control penalty and deterministic selection rules. Both initial language fallback and subsequent strict retry were represented in the supplied scripts.

Quality-control policy and selection. The five principal flags in the supplied scripts were: empty transcript; at least five consecutively repeated words; at least four occurrences of a trigram together with at least 25% repeated-trigram coverage; more than 6.5 words per second when at least eight words were returned; and detection of unexpected non-Latin characters. Corresponding QC penalties were 100 (empty), 50 (repeated word), 40 (repeated phrase), 25 (high word rate), and 80 (non-Latin script), with further penalties for excess repetition. When language-restricted candidates were generated, the candidate with the lower QC penalty was preferred. Ties were resolved by similarity to the first automatic hypothesis and then a deterministic Malay preference. Strict retries of flagged windows selected a new result only when its penalty improved. These rules reduce obviously implausible outputs but may change the evaluation distribution; the source materials did not include the complete raw-versus-final hypothesis pairs needed to quantify this influence.

TABLE I. ASR PIPELINES AND AUDITED SCRIPT-LEVEL DEFAULTS

Pipeline	Checkpoint/inference engine	Audited source-script configuration
1	Mesolitica Malaysian Whisper Large-v3-Turbo-v3; Transformers	Float16; one-beam deterministic generation; max_new_tokens=256; no contextual prompt
2	OpenAI Whisper Large-v3-Turbo; faster-whisper/CTranslate2	Default float16; beam_size=1, best_of=1, temperature=0; max_new_tokens=256; no contextual prompt
3	Qwen/Qwen3-ASR-1.7B-hf; Transformers	Bfloat16 when supported, else float16; do_sample=False; max_new_tokens=512; combined contextual prompt
4	Qwen/Qwen3-ASR-1.7B-hf; Transformers	Same architecture and script settings as Pipeline 3 but contextual prompt disabled

D. Conventional ASR Evaluation

For reproducibility, hypothesis and reference strings were normalized by Unicode NFKC conversion, lowercasing, apostrophe removal, conversion of hyphens and slashes to spaces, removal of punctuation and symbols, and whitespace reduction. The resulting tokens were compared without translation or conversion into formal Malay. Token alignment yielded substitution, deletion, and insertion counts. This alignment quantified lexical changes but was not a linguistically

coded error taxonomy; no claims about the relative frequency of meaning-changing negations, numeral changes, named entities, or hallucinations were made on the basis of token counts alone.

WER was calculated as (substitutions + deletions + insertions) / reference-word count. CER used edit distance over normalized characters after spaces were removed. Aggregate error rates were micro-averaged over the 826 reference words and corresponding characters, rather than averaged across windows. Additional outputs included exact-match rate, non-empty output coverage, and per-window WER and CER. The submitted manuscript used 5,000 window-bootstrap repetitions to obtain 95% confidence intervals for micro-averaged lexical scores; these intervals describe sampling variability within the selected recording and do not account for independent recording-to-recording heterogeneity. The independently executable statistical-analysis script accompanies the revision, but reproducing the upstream ASR generation exactly would require the original audio, runtime environment, checkpoint revisions, and raw quality-control logs.

E. Human Evaluation of Semantic Preservation

Human semantic assessment used the same 30 selected windows as the conventional lexical evaluation. Each reference excerpt appeared with four anonymized candidate transcripts labelled A–D, retaining original window identifiers for data traceability and sequential Items 1–30 in the rating form. The ratings therefore compared text hypotheses with the reference transcripts on a matched-item basis. The provided survey workbook did not document separate access to the source audio during rating or independent verification of the reference transcripts.

Three evaluators independently assigned each candidate one of the form's original binary labels: 'Same essential meaning' or 'Different meaning.' They reported self-assessed familiarity with Malay, English, and Malaysian Malay-English code-switching of 4–5 on five-point scales. Analytically, the 'same' category was interpreted as retention of the utterance's principal communicative content, but the supplied records do not contain a prespecified written rubric defining how to judge negation, factual details, numeral substitutions, omissions, or uncertainty. This interpretation is therefore a clarification of the recorded endpoint, not a retrospective claim that standardized instructions or training were provided before rating.

The original primary label for each pipeline-window pair was 'same essential meaning' when at least two of three evaluators selected that category. No consensus meeting or formal adjudication took place in the supplied records, and majority labels are not described as adjudicated reference labels. Fleiss' kappa and the proportion of exact three-rater agreement were calculated from the independent ratings [22], with the traditional Landis–Koch descriptors used only as descriptive context [23]. Given the known sensitivity of majority classifications to disagreement, the reliability of each pipeline's semantic endpoint was examined separately.

Data and code traceability. The matched dataset contains 30 reference excerpts, four final pipeline hypotheses per excerpt, and three independent ratings per hypothesis.

An accompanying reproducible Python analysis script and a configuration audit can reproduce the reported lexical and semantic statistics from these tables. For independent re-execution of the source-audio recognition stage, the original recordings, full package manifest, model revision hashes, final command-line flags, and the raw/retried transcript stages would additionally be required. No public archive or permanent repository identifier is asserted here.

F. Statistical Analysis

Because the four pipelines processed the same 30 windows, item-level WER and CER were compared using Friedman tests, with Kendall's W as a within-item effect size. Pairwise two-sided Wilcoxon signed-rank tests were adjusted using Holm's method. Majority-vote semantic proportions were accompanied by Wilson 95% confidence intervals. The four related binary semantic outcomes were compared with Cochran's Q test; paired exact McNemar tests with Holm correction were used for secondary contrasts. As an additional uncertainty analysis, the Pipeline 1 minus Pipeline 4 difference in majority semantic preservation was estimated by resampling complete matched windows 5,000 times (fixed seed 20261014) and taking percentile confidence limits. This window-level interval does not correct for shared-speaker or shared-recording dependence. Statistical significance was considered at $p < 0.05$, with conclusions tempered by the purposive single-recording design and multiple comparisons.

IV. Results

A. Conventional ASR Performance

Table II summarizes results for all 30 matched excerpts. Pipeline 4 had the lowest observed micro-averaged WER, 29.18% (95% CI 19.52–40.17%), and CER, 20.60% (95% CI 13.55–28.71%). Pipeline 1 followed, with WER 30.39% (95% CI 22.74–39.69%) and CER 21.71% (95% CI 15.61–29.08%). Pipeline 3 had WER 32.20% and CER 23.20%, while Pipeline 2 had WER 65.86% and CER 52.69%. Pipeline 2's WER exceeded Pipeline 1's by 35.47 percentage points, an observed contrast between distinct checkpoints and inference implementations rather than an isolated measure of model adaptation. The wide confidence intervals, particularly for Pipeline 4, indicate substantial uncertainty in the lexical estimates.

Friedman tests detected a difference among the four pipelines for item-level WER (chi-square (3) = 15.64, $p = 0.0013$; Kendall's $W = 0.174$) and CER (chi-square (3) = 18.17, $p < 0.001$; Kendall's $W = 0.202$). Holm-adjusted pairwise tests indicated worse WER for Pipeline 2 than for Pipelines 1, 3, and 4; the WER differences among the latter three were not significant. For CER, Pipeline 2 differed from the other three pipelines, and Pipeline 4 had lower CER than Pipeline 3 (adjusted $p = 0.039$). The total reference-token error composition was also different: Pipeline 1 had 122 substitutions, 65 deletions, and 64 insertions, compared with 337 substitutions, 167 deletions, and 40 insertions in Pipeline 2 (Table II). These counts describe lexical edits, not verified categories of semantic distortion.

TABLE II. CONVENTIONAL ASR PERFORMANCE ACROSS THE MATCHED 30 WINDOWS

w	Model	WER, % (95% CI)	CER, % (95% CI)	Exact match, %	Coverage, %	S / D / I
1	Mesolitica Whisper	30.39 (22.74-39.69)	21.71 (15.61-29.08)	0.00	100.00	122 / 65 / 64
2	OpenAI Whisper	65.86 (53.58-78.34)	52.69 (42.02-63.77)	3.33	100.00	337 / 167 / 40
3	Qwen3 + prompt	32.20 (23.03-42.58)	23.20 (15.87-31.81)	3.33	100.00	150 / 49 / 67
4	Qwen3 no prompt	29.18 (19.52-40.17)	20.60 (13.55-28.71)	13.33	100.00	125 / 40 / 76

B. Effect of Contextual Prompting in Qwen3-ASR

For the Qwen3-ASR checkpoint, disabling the single combined contextual prompt was associated with lower observed aggregate WER (32.20% to 29.18%; 3.03 percentage points, or 9.4% relative) and CER (23.20% to 20.60%; 2.60 percentage points, or 11.2% relative). Pipeline 4 had lower per-window WER in 13 windows, Pipeline 3 in 7, and 10 were tied; the Holm-adjusted WER comparison was not significant ($p = 0.351$). For CER, Pipeline 4 performed better in 18 windows, Pipeline 3 in 6, with 6 ties; this comparison remained significant (Holm-adjusted $p = 0.039$). These data do not establish which component of the combined prompt contributed to the observed differences.

C. Human-Rated Semantic Preservation

Table III presents the unadjudicated three-evaluator majority endpoint. Pipeline 1 had the highest observed semantic-preservation proportion, 53.3% (16/30; Wilson 95% CI 36.1–69.8%), followed by Pipeline 3 at 50.0% (15/30; 95% CI 33.2–66.8%), Pipeline 4 at 36.7% (11/30; 95% CI 21.9–54.5%), and Pipeline 2 at 33.3% (10/30; 95% CI 19.2–51.2%). The intervals overlap broadly; these values represent proportions of majority classifications in the selected sample, not estimates of independently adjudicated semantic accuracy.

TABLE III. THREE-EVALUATOR SEMANTIC-PRESERVATION RESULTS

Pipeline	Majority preserved	Rate, % (95% CI)	Fleiss' kappa	Exact 3/3 agreement, %	2-to-1 split, n
1	16/30	53.3 (36.1-69.8)	0.555	66.7	10
2	10/30	33.3 (19.2-51.2)	0.665	76.7	7
3	15/30	50.0 (33.2-66.8)	0.333	50.0	15
4	11/30	36.7 (21.9-54.5)	0.552	66.7	10

Overall Fleiss' kappa was 0.529, with unanimous agreement for 78 of 120 pipeline-window judgments (65.0%). Kappa was 0.555, 0.665, 0.333, and 0.552 for Pipelines 1–4, respectively. Pipeline 3 had a two-to-one split on 15 of 30 windows, indicating greater rater dependence of its majority result. Cochran's Q did not reject equality of semantic proportions across pipelines ($Q = 5.57$, 3 degrees of freedom, $p = 0.134$), and no pairwise exact McNemar comparison remained significant after Holm adjustment. Pipeline 1 exceeded Pipeline 4 by 16.7 percentage points, but the matched-window bootstrap 95% CI ranged from −6.7 to 40.0 percentage points and therefore included zero.

D. Qualitative Evaluator Observation

Free-text comments from one evaluator mentioned numeral and number-word confusion, replacement of lexical content by fillers, question or punctuation handling, short- and long-window errors, and acoustically similar substitutions that appeared to alter meaning. These observations were not obtained through a prespecified taxonomy or independent audio-supported adjudication. They are included as qualitative indications of possible error mechanisms, not as verified frequency estimates or evidence that a particular error class caused the observed between-pipeline ranking differences.

E. Divergence Between Lexical and Semantic Rankings

The descriptive rankings differed between the two outcome families on the same 30 items (Table IV). Pipeline 4 ranked first by WER and CER but third by majority semantic preservation, while Pipeline 1 ranked second lexically but first for the majority endpoint. Pipeline 3 ranked third lexically and second semantically; Pipeline 2 ranked fourth under both approaches. This is a descriptive mismatch in rankings, not a statistically confirmed reversal in semantic performance: the overall Cochran's Q test and Holm-adjusted pairwise semantic contrasts were nonsignificant, and the Pipeline 1–Pipeline 4 paired confidence interval crossed zero.

TABLE IV. DESCRIPTIVE RANKING BY LEXICAL AND SEMANTIC OUTCOMES

Pipeline	WER, %	WER rank	Semantic preservation, %	Semantic rank
1	30.39	2	53.3	1
2	65.86	4	33.3	4
3	32.20	3	50.0	2
4	29.18	1	36.7	3

V. DISCUSSION

On the selected excerpts, lexical error rates and independent human judgments produced different descriptive rankings. Qwen3-ASR without the combined prompt had the lowest observed WER and CER, while the Mesolitica pipeline had the highest observed rate of majority-assessed semantic preservation. However, the semantic differences were not statistically significant in paired analyses. The present findings therefore illustrate a possible mismatch between lexical and semantic evaluation in one case study rather than demonstrating superior meaning preservation by any pipeline. This interpretation is compatible with earlier work distinguishing token-level restoration from information preservation [14], [15] and with published semantic-sensitive ASR evaluation methods [16]–[20], although those automated metrics were not applied to these data.

WER and CER remain essential, interpretable indicators of transcription quality. The significant overall Friedman tests and the comparatively high lexical errors for Pipeline 2 show that

these outcomes distinguish some complete inference pipelines within the evaluated sample. Yet a token substitution that changes negation, quantity, or a key entity can have a different downstream consequence from a change in filler words or orthography, even when both count as one edit. The present ratings were intended to capture this distinction, but neither a validated severity scale nor audio-verified error taxonomy was available to establish specific causal pathways. Accordingly, the added value of semantic assessment remains a testable proposition rather than a conclusive finding of this small study.

The practical motivation is especially relevant to code-switched speech, for which pronunciation, borrowing, informal syntax, and spelling variation can affect lexical scoring [8], [9]. Prior Malay-English research has documented gains in recognition performance from code-switching adaptation [11], [12]. The current dataset does not establish whether a semantic endpoint predicts user comprehension or performance on summarization, information extraction, or clinical decision support. Prospective downstream evaluation would be necessary before recommending a particular pipeline for such applications.

The Qwen comparison isolated one model-level configuration difference—presence or absence of a single combined contextual prompt—while retaining the same checkpoint and broad generation workflow. The unprompted condition showed lower WER and CER in aggregate, but the adjusted WER contrast was not statistically significant. Because the prompt simultaneously specified language context, code-switching, verbatim transcription, translation avoidance, and script, it was not possible to attribute the results to any one instruction. Prompt composition and potential interactions should be examined experimentally before broader conclusions about prompting are drawn.

The Mesolitica and OpenAI Whisper systems differed in both checkpoint and inference backend. The former used a Malaysia-adapted Whisper checkpoint in Transformers, whereas the latter used a distinct Whisper checkpoint in faster-whisper/CTranslate2. Consequently, the 35.47-percentage-point WER difference does not reveal whether adaptation, software implementation, decoding behavior, model conversion, or their interactions explain the result. Prior use of a Mesolitica-family checkpoint in Malay speech research [7] supports its relevance as a comparator but does not resolve these confounding factors. Any model-adaptation interpretation requires common-backend experiments with matched decoding settings.

The semantic endpoint was also subject to measurement uncertainty. Fleiss' kappa of 0.529 indicates moderate overall agreement according to commonly used descriptive categories [23], but the lower kappa of 0.333 and 15 split judgments for Pipeline 3 show how sensitive a three-person majority can be to evaluator differences. Binary ratings further conceal whether disagreement concerned an omitted factual detail, a partially preserved proposition, ambiguous reference wording, or different judgments of acceptable paraphrase. Because the original labels were not adjudicated and reference-transcript provenance was not independently documented, the ratings should be interpreted as exploratory human judgments rather than verified ground-truth semantic classifications.

Limitations. First, the dataset comprised only 30 purposively selected windows from one recording; shared recording, topic, speaker, and acoustic characteristics restrict external validity, and window-level confidence intervals do not capture recording-level variability. Second, the source of the reference transcripts and any independent audio verification were not documented in the available materials. The survey recorded binary judgments without a preserved prespecified annotation rubric, rater training protocol, disagreement adjudication, or semantic-severity scale; the low agreement for Pipeline 3 further limits certainty. Third, although segmentation was matched, Pipeline 1 and Pipeline 2 used different checkpoints and backends, and other generation settings were not completely harmonized; this design cannot isolate adaptation or inference-software effects. Fourth, the supplied data contained final QC-selected transcripts but not paired pre-QC and post-QC outputs for each window, preventing estimation of selection or retry bias. Fifth, no audio-verified error taxonomy, automated Semantic Distance/Aligned Semantic Distance/SeMaScore analysis, multi-component prompt ablation, downstream task evaluation, or standardized speed and memory benchmarking was performed. Installed software versions, model commit revisions, and the original execution hardware were not fully archived, limiting exact audio-to-transcript reproducibility. These omissions are methodological limitations rather than negative experimental findings.

Future work. The priority is independent validation using multiple recordings and speakers, with prospectively specified stratification by recording, speaker, duration, language-switch density, background noise, and overlapping speech; uncertainty should be assessed by recording- or speaker-level resampling rather than treating all windows as independent. The reference transcription process should be documented, including independent audio verification and a written bilingual semantic rubric specifying treatment of negation, numerals, named entities, omissions, factual relations, paraphrases, and unresolvable cases. Additional blinded raters and an auditable adjudication procedure should be used to produce an adjudicated secondary endpoint without altering the original ratings. Second, an audio-supported error taxonomy should classify substitution, deletion, insertion, quantity, negation, named-entity, code-switching, filler, and hallucination events, with independently assessed impact on meaning and inter-coder agreement. Validated implementations of Semantic Distance, Aligned Semantic Distance, and SeMaScore should then be applied to the same reference-hypothesis pairs and compared with human judgments using appropriate paired or clustered uncertainty methods. Third, a controlled backend-by-checkpoint comparison of compatible Whisper configurations and a Qwen prompt ablation separating language, code-switching, verbatim, and script instructions should be conducted with fixed generation settings and archived raw and post-QC hypotheses. Finally, standardized inference-latency, real-time factor, throughput, GPU/CPU memory, and total preprocessing/QC overhead measurements are required for deployment comparisons, ideally followed by tests of summarization, information extraction, and other downstream outcomes sensitive to meaning.

VI. CONCLUSION

In this exploratory 30-window, single-recording comparison, the Qwen3-ASR pipeline without a combined contextual prompt obtained the lowest observed WER and CER, whereas the Mesolitica pipeline had the highest observed proportion of three-evaluator majority judgments for essential-meaning preservation. Their ordering differed descriptively, but the overall semantic comparison was not significant (Cochran's Q $p = 0.134$); pairwise semantic tests were nonsignificant after multiple-comparison adjustment, and the paired confidence interval for Pipeline 1 versus Pipeline 4 included zero. No claim of semantic superiority, generalizable pipeline ranking, or a causal effect of model adaptation is justified by these data.

These findings motivate, but do not validate, the routine combination of lexical and semantic endpoints when ASR outputs will support meaning-sensitive tasks. Larger independent corpora, prospectively defined and adjudicated annotation procedures, controlled inference comparisons, and operational benchmarking are needed to determine whether human or automated semantic measures improve ASR selection beyond WER and CER.

Data and Code Availability. The anonymized item-level transcript pairs, independent semantic ratings, evaluation script, and audited V5 inference scripts are available in the accompanying revision materials for author review. Any public release of excerpts or source recordings is subject to confirmation of consent, access rights, and data-sharing approval.

REFERENCES

[1] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, “Automatic speech recognition for under-resourced languages: A survey,” Speech Communication, vol. 56, pp. 85-100, 2014, doi: 10.1016/j.specom.2013.07.008.

[2] K. Fatehi, M. T. Torres, and A. Kucukyilmaz, “An overview of high-resource automatic speech recognition methods and their empirical evaluation in low-resource environments,” Speech Communication, vol. 167, Art. no. 103151, 2025, doi: 10.1016/j.specom.2024.103151.

[3] A. Babu, C. Wang, A. Tjandra, et al., “XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale,” in Proc. Interspeech 2022, pp. 2278-2282, 2022, doi: 10.21437/Interspeech.2022-143.

[4] A. Conneau, M. Ma, S. Khanuja, et al., “FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech,” in 2022 IEEE Spoken Language Technology Workshop (SLT), pp. 798-805, 2023, doi: 10.1109/SLT54892.2023.10023141.

[5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” in Proc. 40th Int. Conf. Machine Learning (ICML), PMLR, vol. 202, pp. 28492-28518, 2023, doi: 10.48550/arXiv.2212.04356 (arXiv version).

[6] X. Shi, X. Wang, Z. Guo, et al., “Qwen3-ASR Technical Report,” arXiv preprint arXiv:2601.21337, 2026, doi: 10.48550/arXiv.2601.21337.

[7] B. Zhang, N. A. A. Latiff, J. Kan, R. Tong, D. Soh, X. Miao, and I. McLoughlin, “Automated evaluation of children's speech fluency for low-resource languages,” in Proc. Interspeech 2025, pp. 1948-1952, 2025, doi: 10.21437/Interspeech.2025-1358.

[8] M. B. Mustafa, M. A. Yusoof, H. K. Khalaf, A. A. R. M. Abushariah, M. L. M. Kiah, H. N. Ting, and S. Muthaiyah, “Code-Switching in Automatic Speech Recognition: The Issues and Future Directions,” Applied Sciences, vol. 12, no. 19, Art. no. 9541, 2022, doi: 10.3390/app12199541.

[9] M. T. Agro, A. A. Kulkarni, K. Kadaoui, Z. Talat, and H. Aldarmaki, "Code-Switching in End-to-End Automatic Speech Recognition: A Systematic Literature Review," in Proc. Fifteenth Language Resources and Evaluation Conference (LREC 2026), pp. 9790-9812, 2026, doi: 10.63317/477upr9ikf9n.

[10] A. Biswas, E. Yilmaz, E. van der Westhuizen, F. de Wet, and T. Niesler, "Code-switched automatic speech recognition in five South African languages," Computer Speech & Language, vol. 71, Art. no. 101262, 2022, doi: 10.1016/j.csl.2021.101262.

[11] T. Nguyen and T. H. Dat, “Can we train ASR systems on code-switching without real code-switch data? Case study for Singapore's languages,” in Proc. Interspeech 2025, pp. 753-757, 2025, doi: 10.21437/Interspeech.2025-654.

[12] J. T. Nguyen and H.-D. Tran, "AsyncSwitch: Asynchronous Text-Speech Adaptation for Code-Switched ASR," in 2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 1-7, 2025, doi: 10.1109/ASRU65441.2025.11434656.

[13] C. -K. Yang, K. -P. Huang, K. -H. Lu, C. -Y. Kuan, C. -Y. Hsiao, and H. -Y. Lee, “Investigating Zero-Shot Generalizability on Mandarin-English Code-Switched ASR and Speech-to-Text Translation of Recent Foundation Models with Self-Supervision and Weak Supervision,” in 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), pp. 540-544, 2024, doi: 10.1109/ICASSPW62465.2024.10626762.

[14] 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, pp. 2765-2768, 2004, doi: 10.21437/Interspeech.2004-668.

[15] X. He, L. Deng, and A. Acero, “Why Word Error Rate Is Not a Good Metric for Speech Recognizer Training for the Speech Translation Task?,” in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), pp. 5632-5635, 2011, doi: 10.1109/ICASSP.2011.5947637.

[16] S. Kim, A. Arora, D. Le, C. -F. Yeh, C. Fuegen, O. Kalinli, and M. L. Seltzer, "Semantic Distance: A New Metric for ASR Performance Analysis Towards Spoken Language Understanding," in Proc. Interspeech 2021, pp. 1977-1981, 2021, doi: 10.21437/Interspeech.2021-1929.

[17] S. Kim, D. Le, W. Zheng, T. Singh, A. Arora, X. Zhai, C. Fuegen, O. Kalinli, and M. L. Seltzer, “Evaluating User Perception of Speech Recognition System Quality with Semantic Distance Metric,” in Proc. Interspeech 2022, pp. 3978-3982, 2022, doi: 10.21437/Interspeech.2022-11144.

[18] J. Rugayan, T. Svendsen, and G. Salvi, “Semantically Meaningful Metrics for Norwegian ASR Systems,” in Proc. Interspeech 2022, pp. 2283-2287, 2022, doi: 10.21437/Interspeech.2022-817.

[19] J. Rugayan, G. Salvi, and T. Svendsen, “Perceptual and Task-Oriented Assessment of a Semantic Metric for ASR Evaluation,” in Proc. Interspeech 2023, pp. 2158-2162, 2023, doi: 10.21437/Interspeech.2023-1778.

[20] Z. Sasindran, H. Yelchuri, and T. V. Prabhakar, “SeMaScore: A new evaluation metric for automatic speech recognition tasks,” in Proc. Interspeech 2024, pp. 4558-4562, 2024, doi: 10.21437/Interspeech.2024-2033.

[21] S. Liang, N. Ballier, G. -A. Levow, and R. Wright, “Beyond WER: Probing Whisper's Sub-token Decoder Across Diverse Language Resource Levels,” in Proc. 2025 Conf. Empirical Methods in Natural Language Processing (EMNLP), pp. 31237-31247, 2025, doi: 10.18653/v1/2025.emnlp-main.1591.

[22] J. L. Fleiss, “Measuring nominal scale agreement among many raters,” Psychological Bulletin, vol. 76, no. 5, pp. 378-382, 1971, doi: 10.1037/h0031619.

[23] J. R. Landis and G. G. Koch, “The Measurement of Observer Agreement for Categorical Data,” Biometrics, vol. 33, no. 1, pp. 159-174, 1977, doi: 10.2307/2529310.