

# SUPPLEMENTARY MATERIAL: HALLUCINATION BENCHMARK FOR SPEECH FOUNDATION MODELS

**Anonymous authors**

Paper under double-blind review

## A DATASETS AND MODELS DETAILS

This appendix section provides complete information about the datasets and models used in our experimental evaluation of the SHALLOW benchmark framework. Tables 1 and 2 summarize the key characteristics of the datasets and models, respectively.

Table 1: Summary of datasets used in the SHALLOW benchmark evaluation.

Dataset	# Test Utts	Domain	Characteristics
<i>Standard Speech Conditions</i>			
LibriSpeech (other) Panayotov et al. (2015)	2,939	Read audiobooks	Standard "other" split with more challenging samples
TEDLIUM Hernandez et al. (2018)	1,469	TED talks	Clear, prepared speech by professional speakers
GIGASPEECH Chen et al. (2021)	25,619	Diverse sources	Audiobooks, podcasts, YouTube; diverse topics
<i>Challenging Acoustic Environments</i>			
CHiME-6 Watanabe et al. (2020)	11,027	Dinner parties	Conversational speech with natural domestic noise
<i>Heavily-Accented Domains</i>			
CORAAL Kendall & Farrington (2023)	5,000	Interview speech	Regional varieties of African American Language
CV16-Accented Ardila et al. (2020)	2,197	Crowd-sourced	English utterances with accent variation
GLOBE-v2 Wang et al. (2024)	5,046	Global accents	164 accents from worldwide speakers
SpeechOcean Zhang et al. (2021)	2,500	L2 English	Non-native speakers (L1: Mandarin); children and adults
<i>Specialized Domains and Voices</i>			
MyST Child Pradhan et al. (2024)	13,180	Educational	Children (grades 3-5) with virtual science tutor
VoxPopuli Wang et al. (2021)	1,842	Political speeches	Formal speaking with domain-specific terminology

### A.1 DATASETS

We selected datasets representing diverse speech conditions, domains, and challenges that ASR systems encounter in real-world applications. The following statistics describe the test sets of the respective datasets.

**Standard Speech Conditions:** LibriSpeech (other) Panayotov et al. (2015) contains 2,939 test utterances from read audiobooks that typically yield low WER scores across modern systems. We use the standard "other" split, which includes more challenging speech samples than the "clean" split. TEDLIUM Hernandez et al. (2018) includes 1,469 test utterances from English-language TED talks, representing clear, prepared speech in a presentation setting with professional speakers. GIGASPEECH Chen et al. (2021) comprises 25,619 test utterances from a multi-domain corpus spanning audiobooks, podcasts, and YouTube videos, covering both read and spontaneous speech across diverse topics including arts, science, and sports, with high-quality transcriptions.

**Challenging Acoustic Environments:** CHiME-6 Watanabe et al. (2020) includes 11,027 test utterances recorded during real dinner parties in everyday home environments. This dataset captures conversational speech with natural domestic noise from kitchen appliances, air conditioning, and movement across various room acoustics.

**Heavily-Accented Domains:** CORAAL Kendall & Farrington (2023) contains utterances from the Corpus of Regional African American Language, sampled from sociolinguistic interviews representing regional varieties of African American Language. It includes audio recordings with time-aligned transcriptions. We selected a subset of 5,000 test samples. CV16-Accented Ardila et al. (2020) consists of 2,197 test utterances from the CommonVoice corpus, specifically selected as English utterances labeled with accent variation. GLOBE-v2 Wang et al. (2024) provides 5,046 test utterances

with worldwide English accents, covering 164 accents from over 23,000 speakers, making it ideal for testing accent generalization. SpeechOcean Zhang et al. (2021) includes 2,500 test utterances from non-native English speakers whose first language is Mandarin, with balanced data from both children and adults with expert-scored pronunciations.

**Specialized Domains and Voices:** MyST Child Pradhan et al. (2024) includes 13,180 test utterances with transcription from children in grades 3-5 conversing with a virtual science tutor, combining children’s speech patterns with scientific vocabulary in educational applications. VoxPopuli Wang et al. (2021) contains 1,842 test utterances from political speeches, offering transcribed formal speaking styles with domain-specific terminology.

Table 2: Summary of ASR models evaluated in the SHALLOW benchmark.

Model	Architecture Type	# Params	Key Characteristics
<i>Self-Supervised Speech Encoders</i>			
HuBERT Hsu et al. (2021)	Encoder-only	300M	Masked prediction objectives; fine-tuned on LibriSpeech
MMS Pratap et al. (2024)	Encoder-only	1B	Multilingual (1,406 languages); language-agnostic representations
<i>Encoder-Decoder Transformers</i>			
Whisper-Large-v2 Radford et al. (2023)	Encoder-decoder	1.5B	680,000 hours of weakly supervised multilingual training
Whisper-Large-v3	Encoder-decoder	1.5B	5M+ hours training data; enhanced generalization capabilities
Canary Puvvada et al. (2024)	Encoder-decoder	1B	FastConformer encoder (32 layers); token-driven decoding
<i>Encoder-Transducer Models</i>			
Parakeet Xu et al. (2023)	Encoder-transducer	1.1B	FastConformer-based; optimized for English recognition
<i>Multimodal SpeechLLMs</i>			
SALMONN Tang et al.	Decoder w/ encoders	7B	Integrates LLMs with speech/audio encoders; unified processing
Qwen2Audio Chu et al. (2024)	Decoder w/ encoders	8.4B	Part of Qwen2 series; specialized audio encoders
Qwen2.5-Omni Xu et al. (2025)	Decoder w/ encoders	10.7B	Enhanced Qwen2; broader multimodal capabilities
Granite-Speech Granite Team (2024)	Decoder w/ encoders	8.6B	Two-pass design for transcription and translation
Kimi-Audio Ding et al. (2025)	Decoder w/ encoders	9.7B	Open audio model; unified framework for audio tasks
Phi4-MM-Instruct Abouelenin et al. (2025)	Decoder w/ encoders	5.6B	Open-weights foundation model; Multimodal by design.

## A.2 MODELS

We evaluated representative models from four distinct ASR architecture families, each employing different approaches to speech processing.

**Self-Supervised Speech Encoders:** HuBERT<sup>1</sup> Hsu et al. (2021) is a self-supervised model trained on masked prediction objectives and fine-tuned on 960 hours of LibriSpeech data. It uses discrete speech units learned through iterative clustering and has demonstrated strong performance on several downstream speech tasks. MMS<sup>2</sup> Pratap et al. (2024) is a multilingual speech encoder based on the wav2vec 2.0 architecture Baevski et al. (2020), trained on 1,406 languages. Unlike language-specific models, MMS extracts language-agnostic representations that aim to generalize across linguistic patterns. Encoder-only models typically focus on acoustic fidelity and may struggle in generating linguistically coherent outputs, potentially impacting morphological and semantic hallucination metrics.

**Encoder-Decoder Transformers:** Whisper-Large-v2<sup>3</sup> Radford et al. (2023) is an encoder-decoder transformer trained on 680,000 hours of weakly supervised multilingual data, demonstrating impres-

<sup>1</sup><https://huggingface.co/facebook/hubert-large-ls960-ft>

<sup>2</sup><https://huggingface.co/facebook/mms-1b-all>

<sup>3</sup><https://huggingface.co/openai/whisper-large-v2>

sive zero-shot generalization across diverse domains and acoustic conditions. Whisper-Large-v3<sup>4</sup> is an enhanced version trained on over 5 million hours of data, maintaining the architecture of its predecessor with refinements to enhance generalization capabilities. Canary<sup>5</sup> Puvvada et al. (2024) is a specialized encoder-decoder model with a FastConformer encoder (32 layers) and a transformer decoder (4 layers), comprising approximately 883M parameters. This model uses token-driven decoding for controlling transcription format, timestamps, and multilingual capabilities. Encoder-decoder models balance acoustic and linguistic modeling, potentially showing more controlled hallucination patterns across multiple dimensions compared to other architectural families.

**Encoder-Transducer Models:** Parakeet<sup>6</sup> Xu et al. (2023) is a FastConformer-based encoder-transducer model optimized for English speech recognition. Transducers employ monotonic alignment between audio and text, potentially influencing their hallucination patterns in continuous speech. The joint network creates tighter coupling between acoustic and linguistic components, which may yield distinct hallucination behavior compared to more loosely coupled encoder-decoder systems.

**Multimodal SpeechLLMs:** SALMONN<sup>7</sup> Tang et al. integrates pre-trained text-based LLMs with speech and audio encoders, processing speech, audio events, and music within a unified framework. Qwen2Audio<sup>8</sup> Chu et al. (2024) is part of the Qwen2 series, with the decoder-only LLM processing audio signals through specialized encoders before generating text responses. We also evaluated Qwen2.5-Omni<sup>9</sup> Xu et al. (2025), which support broader multimodal capabilities. Granite-Speech<sup>10</sup> Granite Team (2024) is a compact decoder-only model employing a two-pass design for transcribing and translating audio inputs. Kimi-Audio<sup>11</sup> Ding et al. (2025) is an open audio model supporting a range of audio processing tasks (including ASR) within a single unified framework. Phi4-Multimodal-Instruct<sup>12</sup> Abouelenin et al. (2025) is an open-weights multimodal foundation model that processes speech inputs alongside text and images. It shows state-of-the-art performance on ASR task. Decoder-only models have stronger language modeling capabilities, which may result in more fluent outputs but potentially higher phonetic or lexical hallucinations due to stronger linguistic priors.

All models were evaluated using authors-provided pre-trained weights without domain-specific fine-tuning to assess their intrinsic hallucination characteristics.

## B SYNTHETIC BENCHMARK DATASET

To rigorously evaluate the SHALLOW metrics under controlled conditions, we introduce a synthetic benchmark dataset designed to isolate individual types of hallucination phenomena in ASR transcriptions. This dataset enables precise analysis of how each metric responds to specific, targeted perturbations, which would be difficult to disentangle in naturally occurring ASR errors.

### B.1 MOTIVATION

While real-world speech corpora are essential for measuring end-to-end ASR performance, they often contain entangled sources of error, i.e., acoustic noise, disfluencies, dialectal variation, and domain mismatch, making it difficult to attribute hallucination metrics to specific error types. In proposing the SHALLOW framework, we wanted to isolate individual hallucination phenomena to validate each metric responds specifically to its intended error category. Aggregate measures like WER offer no insight into the structure of such errors. In contrast, a synthetic dataset allows us to test metric behavior under clean, deliberately controlled conditions where individual hallucination categories are introduced in isolation.

This enables fine-grained stress testing and validation of key metric properties: interpretability, orthogonality, and semantic sensitivity, particularly in edge cases where WER alone fails.

<sup>4</sup><https://huggingface.co/openai/whisper-large-v3>

<sup>5</sup><https://huggingface.co/nvidia/canary-1b-flash>

<sup>6</sup><https://huggingface.co/nvidia/parakeet-rnnt-1.1b>

<sup>7</sup><https://huggingface.co/tsinghua-ee/SALMONN-7B>

<sup>8</sup><https://huggingface.co/Qwen/Qwen2-Audio-7B>

<sup>9</sup><https://huggingface.co/Qwen/Qwen2.5-Omni-7B>

<sup>10</sup><https://huggingface.co/ibm-granite/granite-speech-3.3-8b>

<sup>11</sup><https://huggingface.co/moonshotai/Kimi-Audio-7B-Instruct>

<sup>12</sup><https://huggingface.co/microsoft/Phi-4-multimodal-instruct>

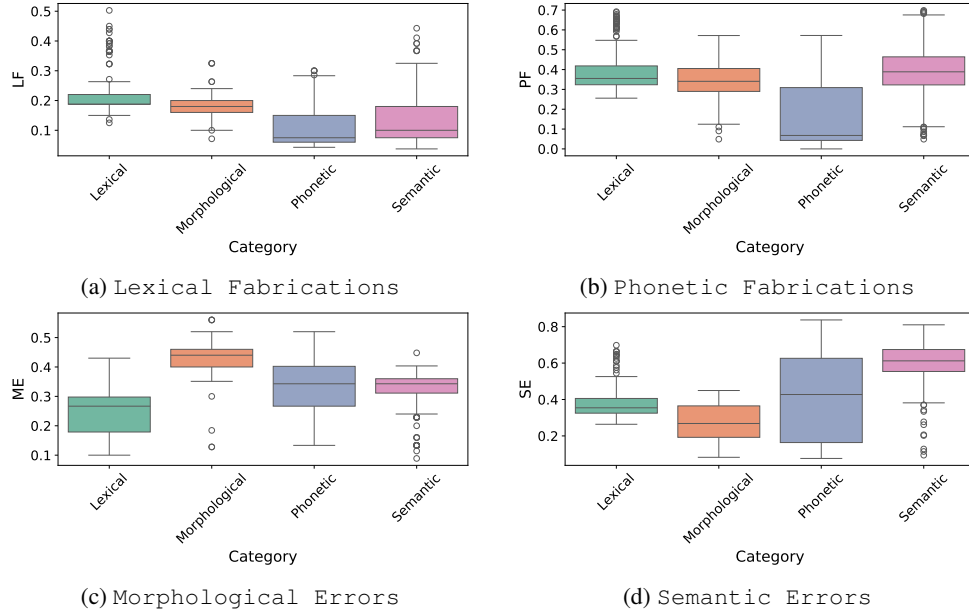


Figure 1: Distribution of hallucination scores across categories for each SHALLOW metric. Each subplot shows box plots of metric values per hallucination type. For most metrics, scores peak in their intended category. The PF metric is lowest on phonetic samples, reflecting successful detection of phonetic proximity.

## B.2 DATASET COMPOSITION

The dataset consists of 1,050 synthetic hypothesis–reference pairs, evenly distributed across six hallucination categories:

- *Lexical Fabrication (150)*: Fluent hallucinations introducing unrelated content not present in the reference.
- *Phonetic Confusion (150)*: Substitutions involving phonetically similar but incorrect words (e.g., “there” vs. “their”).
- *Morphological Divergence (150)*: Grammatical or punctuation-level distortions (e.g., verb tense, agreement, or sentence boundaries).
- *Semantic Drift (300)*: Shifts in meaning, including polarity reversals or role inversion, while preserving lexical fluency. Includes both local (150) and global (150) variants.
- *WER-only Divergence (150)*: High surface-level WER but semantically equivalent hypotheses (e.g., paraphrased or reordered content).
- *Mixed Errors (150)*: Hypotheses with multiple overlapping hallucination types, reflecting realistic multi-dimensional failures.

Each reference is a short, unambiguous sentence in standard English. Hypotheses are generated using GPT-4o Achiam et al. (2023) under type-specific prompts to maximize the intended error while minimizing confounding factors. This construction supports precise validation of metric performance on the phenomena they are meant to detect.

## B.3 GENERATION METHODOLOGY

We used GPT-4o to generate synthetic hypotheses from handcrafted references, using structured prompting tailored to each hallucination category. For example:

- For phonetic confusions, we employed a metaphone-based similarity filter to replace content words with phonetically similar alternatives.

Table 3: Examples of synthetic data with WER and all SHALLOW metrics. Each block focuses on different error categories.  $r_i$ ,  $r_d$ ,  $r_s$  report the insertion, deletion, and substitution ratio, respectively.  $H_N$ ,  $L_N$ , and  $1 - JW$  indicate the Hamming distance (normalized), the Levenshtein distance (normalized), and the inverse of the Jaro-Winkler similarity.  $SD$  denotes the structural divergence, while  $E_{Gr}$ ,  $E_{Sp}$ , and  $E_{Pu}$  are the grammar, spelling, and punctuation errors, respectively, which sum up to the grammatical errors  $GE$ .  $L_{w1}$ ,  $L_{w2}$ , and  $L_{w3}$  mark the local semantic scores for windows considering 1, 2, and 3 words, respectively.  $SDist$  stands for semantic distance, while  $1 - SC$  indicates the inverse of the semantic coherence.  $LF$ ,  $PF$ ,  $ME$ , and  $SE$  denote the aggregate scores for lexical, phonetic, morphological, and semantic categories, respectively.

	Reference	Hypothesis	WER	$r_i$	$r_d$	$r_s$	LF		
Lexical	She left her keys at home	She forgot her keys	0.50	0.00	0.33	0.17	0.12		
	We watched the sun set at the beach	We screamed the sun set at the beach and danced	0.38	0.20	0.00	0.13	0.14		
	She opened a window	She breached the wall portal to let space in	2.00	0.56	0.00	0.75	0.50		
	Reference	Hypothesis	WER	$H_N$	$L_N$	$1 - JW$	PF		
Phonetic	She bakes with flour	She baks with flower	0.50	0.15	0.07	0.02	0.08		
	I cleaned the kitchen	I leaned the kitchen	0.25	0.83	0.08	0.02	0.31		
	I will buy it for you	Isle by it 4 ewe	0.83	0.93	0.43	0.36	0.57		
	Reference	Hypothesis	WER	$SD$	$E_{Gr}$	$E_{Sp}$	$E_{Pu}$	$GE$	$ME$
Morph.	We enjoy watvching birds	We enjoy watching birds frequentlier	0.25	0.20	0.00	1.00	0.00	0.08	0.13
	He painted the wall red	He paints walls redly	0.80	1.00	0.00	0.00	0.00	0.00	0.40
	They ride horses	They rided horses quickerly	0.67	1.00	2.00	0.00	0.00	0.20	0.52
	Reference	Hypothesis	WER	$L_{w1}$	$L_{w2}$	$L_{w3}$	$SDist$	$1 - SC$	$SE$
Semantic	I picked a red flower	I picked a dead flower	0.20	0.96	0.71	0.54	0.40	0.77	0.46
	The big house is old	The small house is new	0.40	0.94	0.73	0.52	0.64	1.00	0.67
	He played video games	He fought sports	0.75	0.65	0.34	0.18	0.71	1.00	0.77

- For semantic drift, we prompted the model to alter meaning without obvious lexical deviation, ensuring plausibility and fluency.
- Morphological errors were crafted by introducing subject-verb agreement errors or incorrect tenses.
- WER-only examples involved paraphrasing references such that WER increases while meaning is preserved, stressing the metric’s discriminative capacity.

Each pair was manually reviewed to ensure alignment with the intended category and avoid noise from model hallucination or overlap.

#### B.4 EXAMPLES

Table 3 shows some representative examples from the synthetic benchmark, illustrating how different error types are instantiated.

#### B.5 METRIC DISTRIBUTION ON THE SYNTHETIC DATASET

Figure 1 presents the distribution of SHALLOW metric scores across synthetic samples, grouped by their intended hallucination category. Each subplot shows a box plot for one metric (Lexical Fabrications  $LF$ , Phonetic Fabrications  $PF$ , Morphological Errors  $ME$ , Semantic Errors  $SE$ ) computed over the synthetic pairs stratified by hallucination type (Lexical, Morphological, Phonetic, Semantic).

The goal of this analysis is to validate the specificity and discriminative capacity of each SHALLOW metric: ideally, a given metric should produce the highest values for samples in its target category, while assigning relatively low scores to samples from other categories. This behavior would confirm that the metrics are aligned with their intended error modalities and are not conflating unrelated phenomena.

**Lexical Fabrication (a):** As expected, the LF metric exhibits the highest values for samples in the Lexical category, indicating that these hypotheses introduce content absent from the reference. Other categories yield lower scores, with the median sharply reduced, consistent with the dataset’s design to minimize lexical novelty outside the intended axis.

**Phonetic Fabrication (b):** Unlike the other panels, the PF metric shows an inverted pattern: the Phonetic category has the *lowest* median score. This is by design. In this benchmark, phonetic hallucination samples were generated by introducing phonetically plausible substitutions (e.g., “there” → “their”), which should yield low phonetic distance if the metric works correctly. Thus, PF scores being minimized here is a positive validation: it confirms that the metric detects phonetic proximity rather than penalizing substitutions indiscriminately.

**Morphological Errors (c):** The ME metric peaks in the Morphological category, as intended. These errors often involve tense, number, and overall sentence structure (e.g., “The cat run” vs. “The cat runs.”), which are designed to specifically challenge grammatical and structural consistency. Other categories display modest scores, affirming metric specificity.

**Semantic Errors (d):** The SE metric exhibits highest median scores for the Semantic category, capturing both local and global meaning shifts. While samples from other categories may contain some incidental semantic variation, their scores remain clearly lower, validating the semantic isolation in the dataset construction.

Taken together, these distributions empirically confirm that SHALLOW metrics react most strongly to their corresponding hallucination types and remain relatively unaffected by unrelated errors. This demonstrates both the targeted design quality of the synthetic benchmark and the functional separability of SHALLOW metrics, which is crucial for their use in detailed ASR hallucination diagnostics. This synthetic dataset thus plays a fundamental role in validating SHALLOW by allowing: (i) *Metric specificity testing*, ensuring each metric responds only to its target error category; (ii) *Correlation analysis*, demonstrating low inter-metric correlation in isolated conditions; (iii) *Controlled counterexamples*, stress-testing metrics on adversarial or benign WER-only cases. This benchmark is released as part of the SHALLOW framework<sup>13</sup> to facilitate reproducibility, benchmarking, and future research into fine-grained ASR hallucination detection.

## B.6 SPEARMAN CORRELATION ANALYSIS

Figure 2 shows the Spearman correlation matrix computed over the synthetic dataset, assessing the relationships between each SHALLOW metric and WER. As expected, the Lexical Fabrications (LF) metric exhibits an almost perfect correlation with WER ( $\rho = 0.98$ ), confirming that lexical insertions and substitutions are the primary drivers of overall word-level mismatch in most ASR hallucinations. Phonetic Fabrications (PF) and Morphological Errors (ME) show moderate positive correlations with WER ( $\rho = 0.54$  and  $0.51$ , respectively), suggesting that these dimensions contribute to error accumulation but are not always aligned with aggregate WER changes. Semantic Errors (SE) are only weakly correlated with WER ( $\rho = 0.15$ ), reinforcing the idea that semantically misleading outputs can occur even when WER is low, and vice versa. The low correlations between SE and other metrics (e.g., LF–SE:  $0.12$ , ME–SE:  $0.13$ ) further highlight the orthogonality of semantic hallucinations within the SHALLOW framework. These findings support our core claim: SHALLOW captures complementary error dimensions that WER alone fails to distinguish, particularly in cases where fluency masks semantic distortion.



Figure 2: Spearman correlation of hallucination scores, synthetic data.

<sup>13</sup>See: <https://anonymous.4open.science/r/SHALLOW/>

## C METRIC IMPLEMENTATION DETAILS

This section provides implementation-specific details for the SHALLOW metrics described in Section 3. We focus on computational considerations, optimizations, and technical choices that complement the theoretical framework presented in the main paper.

### C.1 LEXICAL FABRICATION METRICS

The lexical fabrication metrics quantify word-level deviations between reference and hypothesis transcripts. We implement these metrics using the `JiWER` library<sup>14</sup> to compute insertions, deletions, and substitutions between transcription pairs. For each reference-hypothesis pair, we calculate the relative ratios of these error types. Insertion ratio is computed as the number of inserted words divided by the total word count in the hypothesis. Deletion ratio represents removed words relative to the reference length. Substitution ratio captures replaced words as a proportion of reference length.

Special handling is implemented for edge cases, including empty references or hypotheses.

```

1: if reference = hypothesis then
2:   return {ins = 0, del = 0, sub = 0}           ▷ Short-circuit for exact matches
3: else if len(reference) = 0 then
4:   return {ins = |hypothesis|, ins_ratio = 1.0, del = 0, sub = 0}
5: else if len(hypothesis) = 0 then
6:   return {ins = 0, del = |reference|, del_ratio = 1.0, sub = 0}
7: end if

```

Our implementation detects and excludes common speech disfluencies (e.g., “um,” “uh”) from the insertion count when applying the final weighting formula, as these are considered standard elements of conversational speech rather than hallucinations.

### C.2 PHONETIC FABRICATION METRICS

Phonetic fabrication metrics evaluate the degree of phonetic dissimilarity between reference and hypothesis transcriptions. Our implementation leverages the `Jellyfish` library<sup>15</sup> to transform textual content into metaphone representations, which normalize pronunciation variations. This phonetic encoding converts words to approximate phonetic equivalents, enabling comparison based on pronunciation rather than spelling. We compute three complementary phonetic distance metrics between the metaphone-encoded reference and hypothesis:

1. *Hamming distance*: Measures character-for-character differences, normalized by the length of the longer string between the reference and hypothesis.
2. *Levenshtein distance*: Quantifies the minimum number of single-character edits (insertions, deletions, substitutions) required to transform one string into another, also normalized by the maximum string length.
3. *Jaro-Winkler similarity*: Captures character transpositions and common prefixes, returning a similarity score between 0 and 1.

All distance metrics are normalized to the  $[0, 1]$  range using the maximum possible distance (i.e., the longer string length) rather than using absolute values, enabling consistent scaling across utterances of different lengths. The combined score (as described in Section 3.2) provides a robust measure of phonetic discrepancy that accounts for different aspects of pronunciation variation.

### C.3 MORPHOLOGICAL ERROR METRICS

Morphological error metrics assess structural and grammatical distortions in ASR outputs. Our implementation combines syntax tree comparison with grammar checking to evaluate how ASR systems preserve linguistic structure.

<sup>14</sup><https://github.com/jitsi/jiwer>

<sup>15</sup><https://github.com/jamesturk/jellyfish>

For structural analysis, we use `SpaCy` Honnibal et al. (2020) with the Berkley neural constituency parser Kitaev & Klein (2018)<sup>16</sup> to build dependency trees for both reference and hypothesis texts. Each sentence is represented as a set of dependency relations in the form of (head, dependency relation, token) triples. We compute structural divergence using the Jaccard distance between the reference and hypothesis dependency sets:

$$SD = 1 - \frac{|R \cap H|}{|R \cup H|} \quad (1)$$

where  $R$  and  $H$  represent the sets of dependency relations for reference and hypothesis, respectively. This metric captures differences in grammatical relationships and word order that may affect interpretation.

For grammatical error analysis, we employ the `LanguageTool` API<sup>17</sup> to detect and categorize errors in the hypothesis text. Errors are classified into three primary categories (e.g., Grammar, Spelling, and Punctuation errors) and aggregated using a specific weighting scheme as described in the main manuscript. The final morphological error score integrates both structural and grammatical error analysis into a final score as described in Section 3.3.

#### C.4 SEMANTIC ERROR METRICS

Semantic error metrics evaluate the preservation of meaning between reference and hypothesis transcriptions. Our implementation distinguishes between local semantic errors (affecting short spans) and global semantic coherence (affecting overall meaning).

For local semantic analysis, we employ a multi-scale sliding window approach using contextual embeddings from BERT-based models Devlin et al. (2019). For each window size  $w \in \{1, 2, 3\}$  (unigrams, bigrams, trigrams), we:

1. Compute contextual embeddings for each window in both the reference and the hypothesis;
2. Compare each hypothesis window to all reference windows of the same size using cosine similarity;
3. Retain the maximum similarity score for each hypothesis window;
4. Average these maximum scores, normalized by the length of the longer sequence.

The local semantic error score is computed using a weighted scheme for different window sizes as described in Section 3.4.

For global semantic analysis, we compute two complementary metrics:

1. *Semantic distance* ( $SDist$ ): Computed as the inverse of cosine similarity between sentence-level embeddings generated by a RoBERTa-based model Liu et al. (2019) optimized for NLI tasks.<sup>18</sup>
2. *Semantic coherence* ( $SC$ ): Combines BERTScore F1 with a contradiction-aware penalty from a BART-based Lewis et al. (2020) natural language inference (NLI) model.<sup>19</sup>

Extending previous work on the importance of the semantic dimension in ASR evaluation Kim et al. (2021), our semantic coherence score integrates NLI predictions by scaling the BERTScore with an entailment probability factor:

- 1.0 for entailment classification (reference entails hypothesis)
- 0.5 for neutral classification (no clear relationship)
- 0.0 for contradiction classification (reference contradicts hypothesis)

The global semantic error score averages these components and the final semantic error score combines local and global components with a 1:3 ratio.

<sup>16</sup><https://github.com/nikitakit/self-attentive-parser>

<sup>17</sup><https://languagetool.org/http-api/>

<sup>18</sup><https://huggingface.co/sentence-transformers/nli-roberta-base-v2>

<sup>19</sup><https://huggingface.co/facebook/bart-large-mnli>



Table 4: WER and SHALLOW metrics evaluated on all datasets. Dataset classes are indicated as: Standard Speech , Challenging Acoustic , Heavily-Accented , Specialized Domains , and AVG (overall average) . Best results per dataset underlined, best results on average in **bold**.

Dataset	Metrics	Models											
		HuB	MMS	W-Lv2	Canary	W-Lv3	Parakeet	SALM.	Q2A	Granite	Kimi	Q2.50	Phi4
CHIME-6	WER	59.41	57.30	32.43	34.16	30.25	<u>29.23</u>	136.93	30.93	41.08	33.59	29.92	29.42
	LF	24.20	24.46	15.16	13.20	14.76	13.80	18.53	11.27	13.84	17.90	13.56	15.3
	PF	53.39	55.66	33.20	32.76	<u>30.89</u>	33.49	38.85	33.40	33.41	42.84	32.92	37.36
	ME	37.32	40.27	18.34	19.00	<u>17.28</u>	20.01	22.10	18.46	18.44	23.30	19.04	21.29
	SE	48.02	51.30	26.88	29.45	<u>25.17</u>	27.78	32.31	27.79	27.15	32.83	25.26	30.43
CORAAL	WER	45.05	52.74	22.85	<u>16.58</u>	19.96	22.47	75.08	27.34	22.56	24.16	22.89	23.67
	LF	15.82	19.32	12.94	7.77	10.20	9.56	12.31	7.49	8.79	12.02	8.31	10.39
	PF	40.57	44.55	28.19	<u>21.24</u>	24.49	25.59	29.14	25.73	25.23	35.22	27.23	30.6
	ME	32.35	37.88	17.11	<u>14.01</u>	14.68	17.33	17.54	16.26	15.22	19.53	16.24	18.14
	SE	36.35	44.30	23.08	<u>18.28</u>	19.78	21.12	23.08	20.07	20.43	25.69	20.63	24.95
CV16-Accent	WER	96.02	18.43	20.56	8.08	11.37	<u>5.71</u>	46.26	90.30	6.28	6.87	6.30	6.52
	LF	29.85	5.70	4.23	2.50	3.11	<u>1.71</u>	10.72	26.3	1.93	2.23	2.02	2.09
	PF	69.06	11.75	10.49	6.63	8.12	<u>4.43</u>	43.42	59.95	5.45	6.10	5.61	5.66
	ME	51.16	18.67	9.97	8.39	8.80	6.13	22.24	38.17	6.04	6.60	<u>6.07</u>	6.73
	SE	79.05	16.39	11.55	8.80	9.52	<u>5.58</u>	39.71	68.30	6.56	6.56	6.40	6.36
GigaSpeech	WER	21.13	22.95	15.52	13.79	13.71	<u>11.37</u>	71.62	11.93	18.85	12.64	12.35	12.39
	LF	10.61	14.58	13.91	5.31	13.41	6.37	12.77	<u>5.26</u>	7.38	13.28	7.06	13.02
	PF	26.31	34.87	31.44	<u>16.12</u>	29.16	16.88	27.47	16.36	17.55	31.77	19.65	30.43
	ME	19.77	24.71	16.53	10.15	15.62	10.55	15.91	<u>10.09</u>	10.69	16.61	11.86	16.31
	SE	21.42	27.88	22.95	13.04	22.28	12.81	22.31	<u>12.32</u>	14.59	23.64	13.67	21.49
GLOBE-v2	WER	96.01	12.66	2.89	3.25	1.57	<u>1.17</u>	3.66	4.92	1.47	2.09	3.28	2.68
	LF	30.13	4.42	0.95	1.17	0.58	<u>0.46</u>	1.27	1.61	0.54	0.74	1.02	1.01
	PF	66.80	9.4	2.89	3.39	2.00	<u>1.24</u>	4.34	6.68	1.94	3.54	4.69	3.52
	ME	52.76	14.23	2.73	3.76	1.96	<u>1.50</u>	4.08	5.19	1.73	2.34	3.68	3.11
	SE	78.55	11.91	2.87	3.84	1.87	<u>1.18</u>	4.34	4.79	1.74	2.25	2.84	3.03
LibriSpeech	WER	3.51	7.95	6.15	3.88	3.98	<u>2.62</u>	4.94	3.98	2.98	2.75	3.46	3.83
	LF	1.29	2.88	2.45	1.48	1.48	<u>1.00</u>	1.89	1.42	1.13	1.16	1.35	1.56
	PF	4.48	8.7	7.87	5.31	5.24	<u>3.54</u>	7.46	5.38	4.46	4.35	5.22	5.47
	ME	5.68	11.44	7.58	5.92	5.55	<u>4.35</u>	7.09	5.71	4.47	4.42	5.33	5.80
	SE	3.51	8.81	7.19	5.02	4.63	<u>2.95</u>	6.60	4.58	3.61	3.40	3.96	4.88
MyST	WER	21.98	28.72	20.3	20.99	19.33	<u>13.38</u>	34.46	18.28	18.29	17.64	20.96	14.31
	LF	9.11	12.09	6.89	6.16	6.78	5.61	7.38	<u>5.33</u>	5.79	7.45	6.52	6.54
	PF	24.84	29.42	20.38	22.72	20.19	<u>17.33</u>	20.28	18.76	17.86	22.95	21.63	18.73
	ME	20.45	25.33	12.37	13.34	12.34	11.65	13.18	12.36	11.54	15.00	13.80	12.30
	SE	19.35	26.6	13.97	19.20	13.84	<u>12.50</u>	15.14	13.58	12.63	14.83	14.28	13.37
SpeechOcean	WER	37.98	47.04	25.37	25.35	21.16	23.90	25.98	15.66	24.70	19.92	13.48	<u>12.88</u>
	LF	12.98	15.45	8.19	8.99	7.46	8.14	7.04	5.02	8.41	6.15	4.43	<u>4.27</u>
	PF	21.37	27.75	16.77	17.24	16.17	16.76	15.29	13.64	16.18	15.21	9.83	<u>9.32</u>
	ME	25.30	32.89	15.28	17.15	14.69	16.67	14.88	12.20	15.3	14.59	10.95	<u>10.92</u>
	SE	31.04	41.11	22.43	24.92	21.31	23.69	20.81	16.86	22.34	18.37	14.26	<u>13.76</u>
TEDLIUM	WER	14.26	17.91	18.22	10.29	10.06	10.17	591.01	9.41	10.24	<u>8.12</u>	9.18	9.13
	LF	7.04	8.59	6.81	5.66	6.28	<u>5.37</u>	61.22	5.40	5.93	5.71	5.61	5.75
	PF	27.34	31.62	24.24	22.96	23.91	<u>22.42</u>	75.70	23.97	23.90	25.99	23.30	25.87
	ME	16.88	20.11	12.00	12.25	<u>11.45</u>	11.87	40.24	12.17	11.85	13.02	12.63	11.50
	SE	25.00	26.23	<u>20.73</u>	22.42	20.75	21.62	61.35	21.95	21.99	21.45	21.32	20.99
VoxPopuli	WER	14.05	8.75	26.92	6.22	10.57	<u>5.42</u>	9.21	7.17	5.65	7.53	5.77	5.86
	LF	4.59	2.77	9.28	2.02	3.30	<u>1.75</u>	2.79	2.15	1.86	2.54	1.83	1.90
	PF	21.48	15.66	28.37	12.99	17.29	<u>11.59</u>	17.04	14.30	12.05	16.52	12.45	12.37
	ME	13.88	9.88	19.54	6.55	8.97	<u>5.80</u>	8.14	7.06	6.02	7.57	5.98	6.09
	SE	10.69	6.54	22.09	4.88	8.22	<u>4.11</u>	6.64	5.25	4.60	5.77	4.47	4.41
AVG	WER	40.94	27.45	19.12	14.26	14.20	12.54	99.92	21.99	15.21	13.53	12.76	<b>12.07</b>
	LF	14.56	11.02	8.08	5.43	6.74	5.38	13.59	7.13	5.56	6.92	<b>5.17</b>	6.18
	PF	35.56	26.94	20.38	16.14	17.75	<b>15.33</b>	27.90	21.82	15.80	20.45	16.25	17.94
	ME	27.56	23.54	13.15	11.05	11.13	10.59	16.54	13.77	<b>10.13</b>	12.29	10.56	11.22
	SE	35.29	26.11	17.37	14.99	14.74	13.33	23.23	19.55	13.56	15.48	<b>12.71</b>	14.37

## D COMPREHENSIVE RESULTS ACROSS SPEECH DATASETS

Table 4 reports WER and the four SHALLOW metrics (Lexical, Phonetic, Morphological, Semantic) for all twelve ASR systems evaluated on the ten speech corpora, as well as their corpus-averaged values. Below, we highlight key patterns that underscore the complementary diagnostic power of SHALLOW beyond WER alone.

### D.1 MODEL-LEVEL TRADE-OFFS

**Encoder-decoder variants** Whisper Large-v2 and Large-v3 demonstrate balanced performance across SHALLOW dimensions, with scores that avoid extreme values in any single category ( $PF \approx 18$ – $20$ ,  $ME \approx 11$ – $13$ ,  $SE \approx 15$ – $17$ ). While their WER scores (19.12% and 14.20%) suggest modest differences in overall accuracy, SHALLOW metrics reveal remarkably consistent error profiles; neither model exhibits the sharp dimensional trade-offs seen in other architectures. This balanced hallucination behavior reflects their encoder-decoder design, which integrates acoustic and linguistic processing without strongly prioritizing either phonetic fidelity or semantic coherence.

**Encoder-transducer models** Parakeet delivers the lowest phonetic fabrication score ( $PF = 15.33$ ) and very competitive morphological, lexical, and semantic error rates. This highlights its architectural strength in jointly optimizing acoustic feature encoding and token prediction, enabling more precise word boundary detection and dependency modeling, which in turn minimizes both surface-level confusions and deeper structural distortions at comparable WER levels.

**Multimodal SpeechLLMs** Phi4 and Qwen2.5Omni achieve very low average WER (12.07% and 12.76%, respectively), yet they do not uniformly minimize hallucination metrics. Phi4, for example, has higher Lexical Fabrication (6.18) and Semantic Error (14.37) than Qwen2.5Omni ( $LF = 5.17$ ,  $SE = 12.71$ ), revealing divergent error profiles despite similar WER. SALMONN presents a different failure pattern: despite being designed as a multimodal SpeechLLM with strong language modeling capabilities, it exhibits catastrophic WER (99.92%) while failing to leverage its architectural advantages; its hallucination scores remain comparable to simpler encoder-only models rather than aligning with the semantic coherence demonstrated by other modern SpeechLLM models. This suggests fundamental transcription failures that prevent the model from utilizing its linguistic capabilities.

### D.2 DATASET-SPECIFIC SENSITIVITIES

**Standard Speech Conditions** On high-quality standard speech corpora, SHALLOW metrics reveal consistent patterns that WER alone cannot capture. For example, in LibriSpeech and TEDLIUM, all systems achieve low WER (3–10%, minus a few exceptions) alongside very low lexical fabrication ( $LF \leq 3\%$  for LibriSpeech,  $\leq 8\%$  for TEDLIUM except for SALMONN), and slightly higher semantic and morphological errors. Phonetic fabrications are instead higher, revealing that even under ideal acoustic conditions, residual phoneme-level confusions remain the primary source of errors, an effect that WER aggregates with other error types and thus obscures.

**Noisy Conversational Speech** On CHiME-6, all models record high phonetic fabrications ( $PF \approx 31$ – $56$ ) and moderate morphological errors ( $ME \approx 18$ – $22$ , except for HuBERT and MMS showing higher values), even when WER varies from 29% (Parakeet) to 137% (SALMONN). This suggests that SHALLOW isolates phonetic breakdown as the primary failure mode under acoustic overlap, a nuance lost if only WER were considered.

**Non-Native and Accented Speech** Accented speech datasets reveal SHALLOW’s diagnostic power in isolating accent-specific challenges that WER alone obscures. On CORAAL, despite WER ranging from 17% to 75%, all models exhibit consistently high PF scores (21–45), indicating that dialectal variation primarily manifests as phonetic confusions rather than lexical fabrications or semantic distortions. This pattern persists even for models achieving reasonable WER, suggesting that accent-induced errors concentrate in the phonetic dimension, a distinction completely invisible to aggregate error metrics. The consistency of elevated PF across architectures, regardless of WER performance, demonstrates how SHALLOW isolates specific failure modes that traditional evaluation

conflates with general transcription quality. Such diagnostic precision enables researchers to target accent robustness improvements at the appropriate architectural level rather than pursuing generic WER gains.

**Child Speech** MyST’s spontaneous child dialogue presents unique challenges: WER rises to 13–34% across models. While lexical fabrications remain relatively low across most models (5–7%, with the exception of HuBERT at 9% and MMS at 12%), morphological (ME  $\approx$  12–25%), semantic errors (SE  $\approx$  12–23%) and phonetic fabrications (PF  $\approx$  17–29%) are substantially higher. These scores reflect disfluencies and non-standard syntax in child speech, which standard acoustic and language models struggle to parse. SHALLOW thus pinpoints that errors here are not just phonetic confusions but genuine structural and meaning distortions.

### D.3 MOTIVATION FOR SHALLOW METRICS

The patterns above demonstrate that:

1. *WER is insufficiently granular*: Models with near-identical WER can have markedly different hallucination profiles (e.g., Phi4 vs. Qwen2.5Omni).
2. *Error modes diverge by dataset*: Noisy or dialectal corpora elevate specific hallucination types (e.g., phonetic in CHiME-6, lexical in CORAAL) that WER alone cannot disentangle.
3. *Architectural trade-offs become visible*: Encoder- and decoder-centric designs show complementary strengths (acoustic vs. linguistic), which SHALLOW quantifies directly.
4. *Semantic hallucinations persist despite low WER*: SHALLOW reveals that meaningful content distortions, especially polarity flips or misattributions, can occur even when overall transcription accuracy appears high (i.e., WER is low).

These observations underscore SHALLOW’s role as a *multi-dimensional* diagnostic toolkit: by decomposing ASR errors into lexical, phonetic, morphological, and semantic axes, it surfaces nuanced failure modes and informs targeted model improvement strategies that aggregate WER cannot provide.

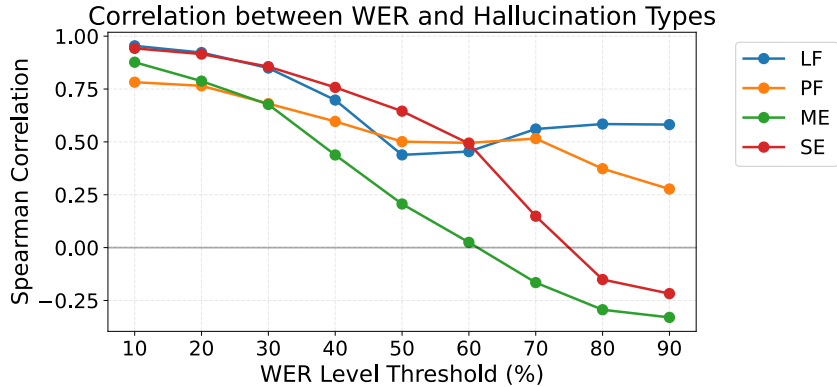


Figure 3: Spearman correlation between WER and each SHALLOW metric, computed on samples filtered by varying WER threshold levels.

## E ADDITIONAL ANALYSIS

### E.1 CORRELATION ACROSS WER THRESHOLDS

Figure 3 presents a threshold-based correlation analysis between WER and the four SHALLOW hallucination metrics: Lexical Fabrication (LF), Phonetic Fabrication (PF), Morphological Errors (ME), and Semantic Errors (SE). We compute Spearman correlation coefficients between WER and each hallucination type, restricting the analysis to model–dataset pairs with WER below increasing thresholds from 10% to 90%.

**Correlation trends.** At low WER levels (below 30–40%), all hallucination metrics are strongly correlated with WER (Spearman  $\rho \geq 0.70$ ), indicating that when models perform well, WER changes largely reflect proportionate reductions in lexical, phonetic, and semantic errors. However, as WER increases, correlations diverge:

- LF remains moderately correlated with WER ( $\rho \approx 0.60$ ) even at high WER, confirming its central role in contributing to raw word errors.
- PF correlation gradually decreases, indicating that phonetic hallucinations become less predictive of WER in degraded conditions.
- ME and SE exhibit sharp correlation drop-offs, eventually turning near-zero or negative (ME:  $\rho < 0$  past 60% WER), showing that morphological and semantic distortion no longer track with raw WER.

These results empirically validate our core claim of the SHALLOW framework: as model performance deteriorates, WER ceases to reliably reflect specific error types, especially those involving meaning and structure, while SHALLOW retains discriminative power.

## E.2 WER-HALLUCINATION CORRELATION HEATMAP

Figure 4 displays a Spearman correlation heatmap between WER and each SHALLOW hallucination type with increasing WER thresholds (from 10% to 90%). This visualization complements the trend plot in Figure 3, offering the same underlying information but at a finer-grained, value-specific level. While the line plot emphasizes overall trends in correlation strength, the heatmap makes it easier to inspect exact correlation values across conditions.

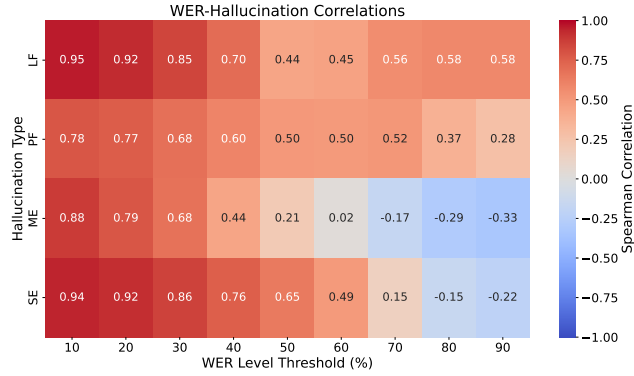


Figure 4: Spearman correlations between WER and each SHALLOW metric, computed over model-dataset pairs with WER below increasing thresholds.

### Strong alignment in low-WER regimes.

At low WER thresholds (10%-30%), all hallucination metrics exhibit strong positive correlations with WER (Spearman  $\rho \geq 0.68$ ), with LF and SE reaching values above 0.90 for lower WER thresholds. This indicates that under high-quality recognition conditions, changes in WER closely reflect changes across all hallucination dimensions, confirming that WER remains a reasonable proxy for error severity when models operate in near-correct regimes.

### Semantic and morphological divergence in higher-WER settings.

As WER thresholds increase beyond 40%, correlations with SE and ME degrade sharply. By 70% WER, the correlation between WER and ME becomes negative ( $\rho = -0.17$ ), and continues decreasing to  $-0.33$  at 90%, indicating that morphological hallucinations become statistically decoupled, and even inversely associated, with WER under severe degradation. Semantic error correlation similarly flips sign beyond 70%, highlighting that meaningful distortions are no longer well-aligned with raw error rate as models deteriorate.

### Lexical and phonetic metrics remain moderately aligned.

In contrast, LF maintains a relatively stable correlation with WER (remaining above  $\rho = 0.44$ ), even at high thresholds. This confirms that lexical fabrication contributes consistently to word-level mismatch across performance ranges. PF exhibits a gradual drop in correlation, settling at  $\rho = 0.28$  at the 90% WER threshold, showing a moderate but diminishing relationship.

### E.3 EXAMPLES

Table 5 shows representative reference-hypothesis pairs from each dataset for six models (Whisper Large-v3, MMS, Parakeet, SALMONN, Qwen2Audio, and Phi-4). These exemplify how WER alone can mask important differences in error types, while SHALLOW metrics reveal the specific nature of hallucinations.

## F COMPUTATIONAL RESOURCES

All SHALLOW experiments and metric evaluations were conducted using a single NVIDIA A100 80GB GPU. This setup was sufficient for both inference over the evaluated ASR systems and full-scale metric computation across all datasets. The complete SHALLOW framework is implemented in a modular and GPU-accelerated fashion where applicable.

**Metric-wise computational complexity.** While WER remains the most widely used metric in ASR evaluation, its simplicity comes with limited diagnostic resolution. SHALLOW metrics provide a richer decomposition of hallucination phenomena, but at the cost of increased computational overhead. Below, we outline the time complexity characteristics per metric:

- *Lexical Fabrication (LF)*: Computed using insertions, deletions, and substitutions derived from Levenshtein alignment. This shares the exact operational backbone with WER and thus incurs negligible additional cost over WER.
- *Phonetic Fabrication (PF)*: Based on phonetic similarity via metaphone, PF is computed per sentence pair and is computationally lightweight, with runtime on par with LF and WER.
- *Morphological Error (ME)*: Involves parsing both hypothesis and reference into dependency graphs using standard syntactic parsers. This step introduces a higher per-sample cost, particularly sensitive to sentence length and syntactic complexity. Runtime grows linearly with the number of tokens and the branching factor of the parse tree.
- *Semantic Error (SE)*: Relies on the computation of sentence-level embeddings (both local and global views), using lightweight transformer-based models. While embedding inference is efficient on modern hardware, SE still incurs a higher cost due to multiple similarity computations (distance and coherence).

**Edge-case robustness.** To prevent unnecessary computation and ensure robustness, SHALLOW incorporates deterministic backoff mechanisms for degenerate cases. If either the reference or hypothesis is empty, or if the pair is exactly equal, metrics are short-circuited to return default values (e.g., zeros or maximum similarity), avoiding meaningless downstream computation.

**Runtime variability.** End-to-end metric computation time varies as a function of (i) Number of samples in the dataset; (ii) Number of edge cases encountered; (iii) Average sentence length per hypothesis–reference pair; (iv) Linguistic complexity (which affects parsing and embedding models); and (v) Number of parallel threads that can be employed. For example, the complete evaluation of the LibriSpeech corpus (3K samples) takes approximately 90 minutes on a single GPU. In contrast, larger and more heterogeneous datasets such as GigaSpeech require more time, depending on batch processing and parser throughput.

**ASR model inference.** Inference for the evaluated ASR systems was conducted using publicly available checkpoints and libraries, all run locally on the same A100 GPU. Models with encoder-only, encoder-decoder, or encoder–transducer architectures (e.g., MMS, Whisper, and Parakeet) exhibit efficient inference times (throughput  $\text{RTFx}^{20} \geq 2300$  for Parakeet), while decoder-only or instruction-tuned SpeechLLMs (e.g., Phi4, SALMONN) show longer inference latencies due to autoregressive decoding (up to 4–5× slower).

<sup>20</sup>Throughput is measured using the RTFx metric, defined as the number of seconds of audio inferred divided by the compute time in seconds. It is the inverse of the RTF (Real Time Factor) metric.

**Scalability and batching.** SHALLOW is designed to process utterances in parallel batches where possible (e.g., embedding-based SE metrics, WER alignment). Parsing-based operations (e.g., ME) remain inherently sequential due to parser design, but can still be parallelized with thread-level concurrency.

SHALLOW incurs modest overhead over traditional WER-based pipelines, especially for metrics requiring linguistic or semantic modeling. Nonetheless, the added interpretability and diagnostic precision justify this cost, especially for applications in critical domains where error type matters more than raw accuracy. Our framework balances efficiency and detail, scaling effectively from small synthetic stress tests to full-scale benchmarks across real-world corpora.

## REFERENCES

- Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benham, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4218–4222, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL <https://aclanthology.org/2020.lrec-1.520/>.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.
- Guoguo Chen et al. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *Interspeech 2021*, pp. 3670–3674, 2021. doi: 10.21437/Interspeech.2021-1965.
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.
- Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.
- IBM Granite Team. Granite 3.0 language models, 2024.
- François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20*, pp. 198–208. Springer, 2018.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020. URL <https://doi.org/10.5281/zenodo.1212303>.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

- Tyler Kendall and Charlie Farrington. The corpus of regional african american language, 2023. URL <https://doi.org/10.7264/1ad5-6t35>. Accessed via The Online Resources for African American Language Project.
- Suyoun Kim, Abhinav Arora, Duc Le, Ching-Feng Yeh, Christian Fuegen, Ozlem Kalinli, and Michael L Seltzer. Semantic distance: A new metric for asr performance analysis towards spoken language understanding. *arXiv preprint arXiv:2104.02138*, 2021.
- Nikita Kitaev and Dan Klein. Constituency parsing with a self-attentive encoder. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2676–2686, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1249. URL <https://aclanthology.org/P18-1249/>.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871. Association for Computational Linguistics, 2020.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.
- Sameer Pradhan, Ronald A. Cole, and Wayne H. Ward. My science tutor (MyST)—a large corpus of children’s conversational speech. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 12040–12045, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.lrec-main.1052/>.
- Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024.
- Krishna C Puvvada, Piotr Żelasko, He Huang, Oleksii Hrinchuk, Nithin Rao Koluguri, Kunal Dhawan, Somshubra Majumdar, Elena Rastorgueva, Zhehuai Chen, Vitaly Lavrukhin, et al. Less is more: Accurate speech recognition & translation without web-scale data. *arXiv preprint arXiv:2406.19674*, 2024.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.
- Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, MA Zejun, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.
- Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 993–1003, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.80. URL <https://aclanthology.org/2021.acl-long.80/>.
- Wenbin Wang, Yang Song, and Sanjay Jha. Globe: A high-quality english corpus with global accents for zero-shot speaker adaptive text-to-speech, 2024.

- Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, David Snyder, Aswin Shanmugam Subramanian, Jan Trmal, Bar Ben Yair, Christoph Boeddeker, Zhaocheng Ni, Yusuke Fujita, Shota Horiguchi, Naoyuki Kanda, Takuya Yoshioka, and Neville Ryant. Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. In *6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, pp. 1–7, 2020. doi: 10.21437/CHiME.2020-1.
- Hainan Xu, Fei Jia, Somshubra Majumdar, He Huang, Shinji Watanabe, and Boris Ginsburg. Efficient sequence transduction by jointly predicting tokens and durations. In *International Conference on Machine Learning*, pp. 38462–38484. PMLR, 2023.
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.
- Junbo Zhang et al. speechocean762: An open-source non-native english speech corpus for pronunciation assessment. In *Proc. Interspeech 2021*, 2021.



Table 5: Examples of evaluated datasets with WER and all SHALLOW metrics.

DS	Model	Hypothesis	Reference	WER	LF	PF	ME	SE
CHIME-6	Wv3	o my god	thank you	1.00	0.27	0.71	0.40	0.67
	MMS		a my gad	0.67	0.20	0.45	0.40	0.37
	Par		o	0.67	0.13	1.00	0.27	0.48
	SALM		i am sorry i did not catch that could you repeat it	4.0	0.68	0.77	0.40	0.62
	Q2A		o my gosh	0.33	0.10	0.20	0.32	0.18
	Phi4		my god	0.33	0.07	0.00	0.13	0.27
CORAL	Wv3	jeremiah he turnt up too	jeremiah you turn to us	0.80	0.24	0.30	0.45	0.61
	MMS		grma ict 0	1.00	0.26	0.60	0.56	0.81
	Par		jeremiah return to	0.80	0.20	0.40	0.48	0.65
	SALM		jeremiah we turn to	0.80	0.22	0.30	0.46	0.49
	Q2A		jeremy yu chang also	1.00	0.28	0.51	0.58	0.37
	Phi4		jeremiah you turned	0.80	0.20	0.30	0.48	0.35
CV16-Accent	Wv3	queuing is something the british excel at	kiwi means something that bridges excel ads	0.71	0.21	0.35	0.37	0.71
	MMS		kiwi knew something that bridgis excel at	0.57	0.17	0.30	0.37	0.54
	Par		queuing is something the british excel at	0.00	0.00	0.00	0.00	0.00
	SALM		kiwi needs something that bridges excel at	0.57	0.17	0.34	0.33	0.54
	Q2A		kids are talking by the door	1.00	0.31	0.58	0.40	0.83
	Phi4		kiwis need something the british excel at	0.29	0.09	0.12	0.27	0.35
GigaSpeech	Wv3	in its hold	and it is old	1.33	0.43	0.49	0.40	0.55
	MMS		in it old	0.67	0.20	0.25	0.40	0.53
	Par		in its hold	0.00	0.00	0.00	0.00	0.00
	SALM		in its hole	0.33	0.10	0.07	0.32	0.66
	Q2A		in its hole	0.33	0.10	0.07	0.32	0.66
	Phi4		ill it hold	0.67	0.20	0.30	0.40	0.47
GLOBE-v2	Wv3	then what does she want with you	yeah i do what does she want to see	0.71	0.24	0.50	0.36	0.53
	MMS		le azil ortas ci wol amfkesi	1.00	0.29	0.69	0.64	0.80
	Par		nadel what does she want with you	0.14	0.04	0.44	0.21	0.14
	SALM		that is all she wants monsieur	0.86	0.24	0.48	0.40	0.65
	Q2A		what does she want pete	0.43	0.10	0.53	0.25	0.45
	Phi4		yeah the other shivaam feature	1.00	0.27	0.76	0.47	0.83
LibriSpeech	WV3	she continued father fauvent	she continued for the fervent .	1.00	0.32	0.27	0.30	0.27
	MMS		she continued father	0.25	0.05	0.23	0.33	0.20
	Par		she continued father fauven	0.25	0.08	0.05	0.33	0.09
	SALM		she continued further prevent	0.50	0.15	0.24	0.27	0.25
	Q2A		she continued father frovent	0.25	0.08	0.10	0.33	0.10
	Phi4		she continued father prevent	0.25	0.08	0.15	0.27	0.15
MyST	Wv3	because we are because we are learning about	because we have been because learning about learning things .	0.75	0.25	0.48	0.38	0.25
	MMS		because we have been becas arling about loingthings i aar	1.00	0.33	0.47	0.50	0.51
	Par		because we have been because running about living things but that	1.00	0.32	0.53	0.40	0.63
	SALM		because we have been because we have been because we have been because we have been [...] because we have to because learning about doing things but the	24.5	0.63	0.82	0.40	0.34
	Q2A		because we have been because learning about living things but but but	1.00	0.32	0.48	0.38	0.37
	Phi4		because we have been because learning about living things but but but	1.13	0.35	0.56	0.38	0.37
SpeechOcean	wv3	alice give up boxing	and skip that book scene	1.25	0.40	0.60	0.40	0.73
	MMS		aris gave tha buksin	1.00	0.30	0.39	0.58	0.68
	Par		alex gave up boxing	0.50	0.15	0.38	0.46	0.51
	SALM		aris give up boxing	0.25	0.08	0.06	0.22	0.14
	Q2A		let us give up boxing	0.50	0.18	0.47	0.29	0.25
	Phi4		alice gave up boxing	0.25	0.07	0.05	0.46	0.09
TEDLIUM	Wv3	and i can twist that around i am sorry if you are getting queasy look away do not look at the thing	and i can twist that around i am sorry if you are getting queasy look away do not look at the thing	0.00	0.00	0.00	0.00	0.00
	MMS		and i can twist that around i am sorry if you are getting queazy look awaydo not look at thei	0.23	0.06	0.12	0.19	0.08
	Par		and i can twist that around i am sorry i do not if you are getting queasy look away do not look at the thing	0.14	0.06	0.25	0.12	0.22
	SALM		thank you for tuning in to our radio show today we are going to be discussing the effects of marijuana on the brain [...]	5.64	0.66	0.76	0.41	0.59
	Q2A		and i can twist that around i am sorry i if you are getting queazy look away do not lok at te thing	0.18	0.06	0.20	0.27	0.09
	Phi4		and i can twist that around i am sorry if you are getting queasy look away do not look at the	0.05	0.01	0.04	0.11	0.04
VoxPopuli	Wv3	i appreciate very much what you said but can you make sure that once you foresee this kind of simulation today that you invite some of the people who were actually in mumbai because it could give you some insight	okay	1.00	0.20	0.83	0.40	0.84
	MMS		ie very much what you said but can you make sure once you foresee this kind of simulation todays that you invite some of the people which were actually in mumbay i think it could be given you some insid	0.28	0.09	0.38	0.28	0.13
	Par		i appreciate very much what you said but can you make sure once you foresee this kind of simulation 2 days that you invite some of the people which were actually in mumbai i think it could give you some insight	0.15	0.05	0.21	0.16	0.24
	SALM		appreciate very much what you said but can you make sure once you foresee this kind of simulation to days that you invite some of the people which were actually in mumbai i think it could give us some insight	0.20	0.07	0.38	0.15	0.16
	Q2A		i appreciate very much what you said but can you make sure once you foresee this kind of simulation to days that you invite some of the people who were actually in mumbai i think it could give you some insight	0.13	0.04	0.28	0.12	0.19
	Phi4		but can you make sure once you foresee this kind of simulation today that you invite some of the people which were actually in mumbai i think it could give you some insight	0.28	0.07	0.45	0.20	0.11