Lost in Transcription, Found in Distribution Shift: Demystifying Hallucination in Speech Foundation Models

Anonymous ACL submission

Abstract

Speech foundation models trained at a massive scale, both in terms of model and data size, result in robust systems capable of performing multiple speech tasks, including automatic speech recognition (ASR). These models transcend language and domain barriers, yet effectively measuring their performance remains a challenge. Traditional metrics like word error rate (WER) and character error rate (CER) are commonly used to evaluate ASR performance but often fail to reflect transcription quality in critical contexts, particularly when detecting fabricated outputs. This phenomenon, known as hallucination, is especially concerning in high-stakes domains such as healthcare, legal, and aviation, where errors can have severe consequences. In our work, we address this gap by investigating hallucination in ASR models. We examine how factors such as distribution shifts, model size, and model architecture influence the hallucination error rate (HER), a metric we introduce to quantify hallucinations. Our analysis of 20 ASR models reveals three key insights: (1) High WERs can mask low hallucination rates, while low WERs may conceal dangerous hallucinations. (2) Synthetic noise, both adversarial and common perturbations like white noise, pitch shift, and time stretching, increase HER. (3) Distribution shift correlates strongly with HER ($\alpha = 0.91$). Our findings highlight the importance of incorporating HER alongside traditional metrics like WER to better assess ASR model performance, particularly in high-stakes domains.

1 Introduction

Automatic Speech Recognition (ASR) systems have become fundamental to various applications, including personal assistants, automated customer service, and transcription tools used in fields such as education, healthcare, and law (Zhang et al., 2023; Adedeji et al., 2024). These systems have seen remarkable improvements in recent years (Arriaga et al., 2024; Radford et al., 2022; Communication et al., 2023), with state-of-the-art models demonstrating their capabilities across diverse datasets and languages (Shakhadri et al., 2025). However, the evaluation of ASR performance remains largely dependent on word and character error rate (WER/CER). The primary limitation of WER and CER is their dependence on token-level overlapping, which focuses on matching individual words or characters without considering the overall semantic aspect of the transcription. This could result in misleading evaluations, as a high WER/CER does not necessarily indicate poor outputs in all cases. In addition, these metrics fall short in capturing more subtle semantic failures which aren't typically caught without human verification, such as hallucinations.

Hallucinations in ASR systems mirror perceptual experiences in neuroscience—plausible outputs generated without grounding in input stimuli (American Psychiatric Association et al., 2013; Zmigrod et al., 2016), deviating *phonetically* or *semantically* from source speech (Ji et al., 2023). Like natural neural perceptions, ASR hallucinations arise when models prioritize distributional patterns over fidelity to audio input, fabricating text unlinked to reference content (Hare, 2021). These errors are uniquely hazardous in high-stakes domains (Williamson and Prybutok, 2024), as they evade WER/CER detection while distorting meaning, similar to how clinical hallucinations disconnect from reality.

Hallucination in domains such as medical and legal can have serious consequences, including life-threatening outcomes and distorted testimonies or contracts, and may disproportionately affect marginalized groups (Xie et al., 2023; Vishwanath et al., 2024; Mujtaba et al., 2024; Sperber et al., 2020; Koenecke et al., 2024a). Hence, detecting and mitigating hallucination is crucial for ensuring the reliability of ASR systems in sensitive environments.

The existing literature on hallucination detection in ASR models is confined to a single model (Frieske and Shi, 2024; Serai et al., 2021; Koenecke et al., 2024a; Barański et al., 2025; Ji et al., 2023) or test setting (Kim et al., 2024), highlighting a significant research gap for a systematic investigation across different supervision paradigms, test domains, and conditions.

In this work, we address this critical research gap and make the following key contributions:

- We conduct a thorough evaluation of ASR models across various setups, consisting of both synthetic and natural shifts from training to test distributions.
- We introduce an LLM-based error detection framework that classifies ASR outputs into different types of errors including hallucination errors through context-aware assessments.
- We provide an in-depth analysis of hallucination phenomena in ASR models, exploring the impact of domain-specific data, model architectures, and training paradigms, and offer valuable insights into the relationship between model size, type, and hallucination frequency.
- To validate our hallucination detection method, we compare our LLM-based hallucination detection pipeline with human evaluations and heuristic approaches, demonstrating that the LLM evaluation closely aligns with human judgments and other LLM-based assessments, unlike the heuristic-based approach.

The significance of this work lies in how it redefines the evaluation and improvement of ASR systems. By emphasizing hallucination detection, we aim to enhance the reliability of ASR models, particularly in domains where accuracy and precision are non-negotiable.

Outline. In Section 2, we review prior work related to ASR evaluation and hallucination detection. Section 3 outlines our proposed methodology. Section 4 presents our experimental setup. Finally, Section 5 discusses the results and implications of our findings, outlining directions for future research. We conclude our work in 6 and provide limitations in 7.

2 Related Work

The use of ASR systems in high-stakes domains, including healthcare (Afonja et al., 2024; Huh et al., 2023; Adedeji et al., 2024; Sunder et al., 2022), legal proceedings (Saadany et al., 2022; Garneau and Bolduc, 2024), and finance (Del Rio et al., 2021; Liao et al., 2023), has heightened the necessity for ensuring their robustness. Conventionally, the performance of these models is assessed using metrics such as WER and CER (Serai et al., 2022). Szymański et al. (2023); Sasindran et al. (2024) They show that when these metrics are used in isolation, they exhibit notable limitations.

Recent studies have extensively investigated hallucination in text generated by large language models (LLMs), identifying it as a prevalent phenomenon (Huang et al., 2025; Bai et al., 2024; Yao et al., 2023; Jiang et al., 2024; Maynez et al., 2020; Parikh et al., 2020; Ji et al., 2023; Mittal et al., 2024; Filippova, 2020). This issue has also been observed in audio foundation models (Sahoo et al., 2024). Furthermore, research suggests that pretraining language models for predictive accuracy inherently predispose them to hallucination, even under ideal conditions where the training data is entirely factual (Kalai and Vempala, 2024).

However, few studies explore hallucination evaluation and detection in automatic speech recognition (ASR) systems, with most research focusing on *Whisper*, a semi-supervised model. For instance, Koenecke et al. (2024b) analyze the Aphasia dataset and report that while *Whisper*'s overall hallucination rate is 1%, 40% of these hallucinations contain violent or harmful content. Similarly, Kim et al. (2024) demonstrate that *Whisper* hallucinates at significantly higher rates under low signal-to-noise ratio (SNR) conditions, and observe a 20% increase in hallucinations at -4 dB and -2 dB SNRs. Prior work by Serai et al. (2021) propose augmenting models with hallucinated transcripts to improve performance, while Frieske and Shi (2024) introduce a perturbation-based evaluation method using automatic metrics such as word error rate (WER), perplexity, and cosine similarity. Barański et al. (2025) develop a filtered Bag of Hallucinations (BoH) for detection, and reveal that hallucinations in *Whisper* correlate strongly with training data biases (e.g., phrases like "Thank you for watching" linked to YouTube content).

Despite these advances, existing studies remain limited in scope, focusing predominantly on semi-

Model	Reference	Hypothesis	WER	Hallucination			Setting
				Heuristic	Human	LLM	
whisper-small	hungry action hippos fruit	Humm reaction in hippos fruit?	75	F	F	F	Home Environment
whisper-large-v3	ripped ocean jumper.	Thank you.	100	F	T	T	Home Environment
hf-seamless-m4t-large	probably i i had asthma	What about Erasmus?	100	F	T	T	Medical
hf-seamless-m4t-large	it can't be done	I'm going to start with the first one.	180	F	T	T	Legal
hf-seamless-m4t-medium	News of this album coincides with John Robert's departure from the band.	The first is the oldest, which is also the oldest, the oldest, the oldest, the oldest, the oldest.	130	F	T	T	Accented Speech
whisper-medium	patel para thirty eight page three hundred and fifty five	How much is the tail?	100	F	T	T	Legal

Table 1: Examples showing hallucination detection by different methods across domains and models.

supervised models like *Whisper*. This highlights a critical gap: the lack of a comprehensive understanding of hallucinations across the full spectrum of supervision paradigms—from supervised to unsupervised models—and under domain shifts where test data distributions diverge sharply from training environments. Addressing these gaps is essential for developing robust ASR systems that maintain accuracy and faithfulness in diverse real-world applications, a challenge our work directly tackles by evaluating a wide range of models with diverse architectures, sizes, and training paradigms on synthetic and natural shifts.

3 Methodology

We examine transcription quality using the standard metrics WER and CER, and we assess the occurrence of hallucination errors to provide a comprehensive view of model performance. Our testing environment is characterized by both natural and synthetic distribution shifts. Furthermore, we investigate the deterioration of error rates when transitioning from a controlled source domain (LibriSpeech) to various target domains, with a particular focus on both WER and the Hallucination Error Rate (HER). In addition, we introduce noise to the input data and analyze its effect on error rate degradation. This multifaceted strategy offers valuable insights into the challenges encountered by ASR systems in real-world scenarios. We provide additional details about each step subsequently.

3.1 ASR Evaluation

We evaluate a broad range of ASR models under a zero-shot setting, using the default decoding parameters for each model. Standard preprocessing steps are applied prior to calculating the metrics, ensuring consistency in evaluation.

3.2 Hallucination Evaluation

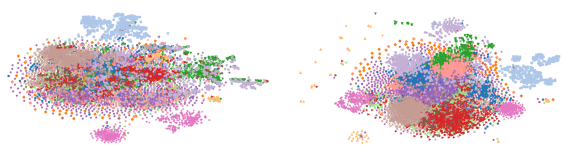
In addition to conventional transcription errors, we also assess hallucination errors by using an LLM-based pipeline that classifies the errors produced by the ASR model. Specifically, we use *GPT-4o mini* to compare the ground truth transcription with the model outputs and ask the LLM to categorize them into different error types.

We conduct hallucination evaluation at two levels:

Coarse-grained: The model categorizes the output into one of three classes: *Hallucination Error*, *Non-Hallucination*, or *No Error*. For this evaluation, we provide two examples per category in the prompt.

Fine-grained: The model is asked to further refine the categorization by identifying specific error types, such as *Hallucination Error*, *Language Error*, *Oscillation Error*, *Phonetic Error*, and *No Error*. In this case, one example per category is provided in the prompt.

The prompts used for both coarse-grained and fine-grained evaluations are detailed in Appendix A.2 (Figure 5 and Figure 6, respectively). To quantify hallucination occurrences, we introduce the *HER*, defined as the ratio of hallucination errors to the total number of examples in the data.



(a) Speech (b) Text

Figure 1: 2D t-SNE representation of our evaluation datasets. The figure (a) shows speech representations, while the figure (b) represents text embeddings, both from SONAR. Each distinct evaluation dataset is represented by unique colors and markers, demonstrating the diversity in both the speech and text of our evaluations.

3.3 Distribution Shifts and Quantification

We systematically evaluate ASR models under a variety of testing conditions, ranging from naturally occurring domain variations to scenarios involving both adversarial and non-adversarial perturbations. These conditions are designed to induce either natural or synthetic distribution shifts in the input speech.

Natural Shifts. These shifts arise from inherent variations in data distributions, such as differences in accents, domain-specific content, background noise, and diverse speaking styles (Liu et al., 2021).

Synthetic Shifts. These shifts are artificially induced, encompassing simulated noise, goal-specific perturbations, and adversarial attacks (Fan et al., 2022). We design a comprehensive testing setup to identify the conditions under which ASR models are prone to hallucinate. Additional details about the datasets used in this study are provided in Section 4.1 and summarized in Table 4. We measure the extent of domain distribution shifts using the high-order metric Central Moment Discrepancy (CMD) (Zellinger et al., 2019; Kashyap et al., 2020), which assesses the discrepancy between two distributions. It is calculated as follows:

$$\text{CMD} = \frac{1}{L}\sum_{l=1}^{L} \left\| E[h_l^s] - E[h_l^t] \right\|_2^2, \quad (1)$$

where L represents the number of layers, h_l^s and h_l^t denote the hidden representations for the source and target domains, respectively, and $E[\cdot]$ signifies the expectation.

3.4 Error Rate Degradation

Error Rate Degradation quantifies the decline in ASR performance when transitioning from a source domain (LibriSpeech) to a target domain, with degradation measured in two aspects: transcription errors and hallucination errors.

Word Error Rate Degradation (WERD). WERD is defined as the difference in WER between the target and source domains:

$$\text{WERD} = \text{WER}_{\text{target}} - \text{WER}_{\text{source}}, \quad (2)$$

where $\text{WER}_{\text{source}}$ and $\text{WER}_{\text{target}}$ denote the WER for the source and target domains, respectively.

Hallucination Error Rate Degradation (HERD). HERD captures the increase in hallucination errors when moving from the source to the target domain:

$$\text{HERD} = \text{HER}_{\text{target}} - \text{HER}_{\text{source}}, \quad (3)$$

where $\text{HER}_{\text{source}}$ and $\text{HER}_{\text{target}}$ represent the hallucination error rates in the source and target domains, respectively.

Furthermore, the relationship between CMD and degradation rates is analyzed and visualized (see Figure 7) to understand how domain variations correlate with both transcription and hallucination errors.

4 Experiments

In this section, we present the experimental details of our work. We examine the models' tendencies to produce hallucinated outputs, using LLM-based evaluation as described in 3. The results provide insights into the reliability of different ASR systems across various real-world and adversarial conditions.

4.1 Datasets

In our experiments, we use a diverse set of datasets representing various domains and testing conditions to evaluate the ASR systems under scenarios that differ from training data. To achieve this, we choose datasets from domains with a high likelihood of being unseen during training, ensuring a natural distributional shift. Additionally, we include datasets with synthetic perturbations, such as adversarial attacks, and common augmentation techniques like Gaussian noise addition, pitch shifting, and time stretching, to assess the model's robustness under synthetic shift.

Domain Specific Datasets. To evaluate model performance under real-world conditions, we leverage datasets from diverse domains, ensuring a comprehensive assessment across various settings. These include legal proceedings: *Supreme-Court-Speech*[1], medical dialogues: *Primock57* (Papadopoulos Korfiatis et al., 2022), meeting conversations: *AMI* (Consortium, n.d.), aviation communications: *ATCOsim* (Hofbauer et al., 2008), conversational speech: *SLUE-VoxCeleb* (Shon et al., 2022), home environments: *BERSt*[2], and general speech corpora: *LibriSpeech* (Panayotov et al., 2015), *GLOBE* (Wang et al., 2024), and *SPGISpeech* (Technologies, 2021), including noisy conditions (*LibriSpeech_test_noise* (Panayotov et al., 2015)). These datasets span a wide

[1] https://huggingface.co/datasets/janaab/supreme-court-speech

[2] https://huggingface.co/datasets/macabdul9/BERSt

range of accents, recording conditions, and environments—from high-quality audiobooks to teleconferences and real-time simulations. This diversity ensures a robust evaluation of model performance across realistic and challenging scenarios, addressing the limitations of single-domain evaluations.

Perturbed Datasets. To simulate challenging acoustic conditions and evaluate WER and HER under adversarial scenarios, we apply various types of synthetic perturbations to speech inputs. These include an adversarial dataset featuring modified utterances with adversarial noise at varying radii (0.04 and 0.015) and Room Impulse Response (RIR) noise, primarily aimed at adversarially attacking the ASR models (Olivier and Raj, 2022). Additionally, we evaluate model robustness under challenging conditions by applying a range of general audio perturbations—including noise addition, time stretching, and pitch shift—to 1,000 randomly sampled audio clips from a mixture of domain-specific datasets. These perturbations are commonly used as augmentation techniques to simulate real-world variability and stress-test the models. We compare the performance on perturbed speech with that of non-perturbed speech to quantify the impact of these distortions. Full details about the perturbation methods, including parameters and implementation, are provided in Appendix A.1.2 (Table 5). Additionally, comprehensive information about the datasets used in this study can be found in Appendix A.1.1 (Table 4).

4.2 Models

To comprehensively evaluate hallucination patterns in ASR systems, we select models that span diverse architectures, sizes, and training paradigms. This diversity enables systematic analysis of how these factors influence hallucination susceptibility. Specifically, we include:

- *Encoder-only* models: HuBERT (Hsu et al., 2021) and Wav2Vec2 (Baevski et al., 2020), which leverage self-supervised training to learn robust speech representations.
- *Decoder-only* models: Qwen2Audio (Chu et al., 2024) and SpeechLLM (Rajaa and Tushar), optimized for text generation and audio-language alignment.
- *Encoder-decoder* models: Whisper (Radford et al., 2022) (10 variants), DistilWhisper (Gandhi et al., 2023) (4 variants), and SeamlessM4T (Communication et al., 2023) (2 variants), designed for multilingual transcription, translation, and speech-to-text tasks.

The selected models vary in size (39M to 7B parameters), depth (4 to 32 layers), and training paradigms (supervised, self-supervised, semi-supervised). Full specifications, including architectural details, training data, and hyperparameters, are provided in Appendix A.1.3 (Table 6).

4.3 Experimental Setup

We utilize models and datasets sourced from Huggingface [3]. All audio data is resampled to match the sampling rate required by the respective models. For each dataset, we randomly sample 1,000 examples from the *test* split to ensure a manageable and consistent experimental setup. Unless otherwise specified, we use the default decoding parameters for ASR evaluation. To analyze the data, we compute SONAR embeddings (Duquenne et al., 2023) for both speech and text. Additionally, we employ CMD based on prior work (Kashyap et al., 2020) to quantify domain shifts. All experiments are conducted on a single A100/H100 GPU. Prior to calculating WER and generating embeddings, we apply a basic English text normalizer [4] to ensure consistency in text preprocessing. For LLM evaluation, we perform greedy search decoding to ensure reproducible outputs.

4.4 Human Evaluation

We construct a human evaluation dataset by aggregating outputs from multiple models, filtering samples with `WER` > 60 to focus on significant deviations. Hypotheses and references are constrained to 1-100 words for balance. To simulate synthetic hallucinations, we shuffle 50 hypotheses, introducing artificial errors (Stiff et al., 2019). The final dataset includes 500 samples, each reviewed by two independent annotators from a pool of 20. This framework ensures robust evaluation and reliable analysis of hallucination patterns across models.

5 Results

In our experiments, we use an LLM-based pipeline to classify ASR errors across various evaluation

[3] `https://huggingface.co/models,https://huggingface.co/datasets`

[4] `https://github.com/huggingface/transformers/blob/main/src/transformers/models/whisper/english_normalizer.py`

setups, validating it against human evaluation and heuristic baselines. We then explore the effects of natural and synthetic distribution shifts on error metrics, specifically examining how domain variations and input perturbations impact word and hallucination error rates. Additionally, we analyze the influence of model architecture and scale through a comparison of Whisper variants and other architectures. This approach provides insights into the complex interactions between error types, data conditions, and model characteristics. In this section, we present our findings.

5.1 Hallucination Error Detection

We assess the ASR outputs of large language models (LLMs) by classifying them into various error categories. Specifically, we use *GPT-4o mini* to identify the types of errors in ASR outputs. The evaluation process includes both coarse-grained and fine-grained error classifications, as detailed in Section 3.2. The prompts used for both evaluations are shown in Figures 5 and 6.

Coarse-grained and Fine-grained Evaluation. Figure 8 in the appendix illustrates the error distributions across coarse-grained and fine-grained hallucination categories. Our results demonstrate strong alignment between both levels, indicating consistent classification of hallucinations. Among non-hallucination errors, phonetic errors dominate across most datasets (see Appendix Table 9). However, in *Primock57*, language errors prevail, likely due to its specialized medical terminology. This aligns with (Ferrando et al., 2024), who emphasize language models' struggles with domain-specific named entities. This is also reflected in the third example provided in Table 1.

Agreement with Human Evaluation and Heuristic Baseline. To validate our approach, we compare *model-to-human* and *human-to-human* agreement scores using a coarse-grained prompt. Our results demonstrate strong human-to-human raw agreement (0.71), indicating consistency. Additionally, we observe good agreement (0.6) between human annotations and *GPT-4o-mini*'s coarse-grained output, suggesting that the model aligns reasonably well with human judgments. We further evaluate the agreement between human and model classifications against a heuristic baseline proposed by (Frieske and Shi, 2024). Their method is based on a cosine similarity threshold of 0.2, alongside a *WER* threshold of 30 and a *Flan-T5* (Chung et al., 2024) perplexity threshold of 200. However, as shown in Table 3, this heuristic achieves significantly lower agreement scores: 0.1 with *GPT-4o mini* and 0.14 with *Gemini-2.0-flash-001*. These results highlight the limitations of purely heuristic-based approaches compared to our method, which better captures the more fine-grained aspects like hallucination.

5.2 Errors Under Distributional Shifts

Natural Shift. Given that most ASR models now outperform the human baseline on the LibriSpeech clean test set, we consider LibriSpeech as the source domain. Other domain-specific datasets, such as Primock, SPGISpeech, GLOBE, and AMI, are therefore treated as the target domain. We compute the distribution shift as detailed in Section 3.3. We then measure the change (degradation) using Equation 2 and 3.

The α is the correlation coefficient between error rate degradation and distribution shift.Figure 2 shows that both WER and HER degrade as we move from the source domain to different target domains, with considerable distribution shifts across various *whisper* models. This degradation exhibits a nearly linear positive correlation with the domain shift.

Notably, the HER demonstrates a slightly stronger correlation with the shift compared to WER. This trend is consistent across all models, as illustrated in Appendix A.3 Figure 7. The *ATCOsim* dataset is an outlier, with artificially high WER due to its numerical content. For example, models generating digits (e.g., "23") instead of spoken forms ("two three") are heavily penalized, inflating WER without accurately reflecting transcription quality.

Synthetic Shift. Under synthetic shift, we experiment with two configurations: (a) adversarial perturbations and (b) common perturbations. Our experiments reveal that adversarial attacks cause the most significant degradation in HER, with adversarial datasets showing the highest HER values across all models, exceeding the degradation observed under natural shift baselines (see Table 2). In contrast, random perturbations (such as white noise, pitch shifts, and time stretching) result in more moderate impacts. Notably, self- and semi-supervised models like *whisper* and *seamless* demonstrate consistent vulnerability to structured perturbations. For instance, pitch shifts and

Model	SPGI	BERSt	ATCOsim	ADV	AMI	SLU	SNIPS	SC	GLOBE	SALT	LS_Noise	LS	Primock57
whisper-large-v3	3.4/0.4	32.4/15.8	65.3/17.6	33.3/49.8	23.4/10.1	15.5/7.9	8.2/0.9	18.8/15.4	3.4/2.5	3.0/1.0	2.6/0.4	2.2/0.3	19.2/4.5
wav2vec2-large-xlsr-53-english	19.6/1.0	64.0/18.6	63.0/19.6	100.1/96.5	53.0/22.7	43.4/6.2	12.4/1.2	32.6/14.8	27.0/10.3	17.0/2.1	9.0/0.6	6.5/0.0	47.9/13.7
hf-seamless-m4t-large	14.7/3.6	58.9/34.6	76.6/62.3	61.2/76.9	63.7/46.2	44.0/23.5	7.4/3.2	34.2/30.0	19.2/21.4	4.2/1.6	6.5/3.2	3.4/0.5	44.5/27.4
speechllm-1.5B	11.5/4.8	68.5/38.5	121.4/57.4	95.3/94.5	127.3/54.4	83.8/16.5	10.8/2.9	41.3/38.4	27.9/28.7	9.5/4.2	10.9/5.2	11.4/4.2	41.7/17.1
whisper-medium	3.7/0.2	34.5/16.2	65.6/18.2	42.9/63.1	23.2/12.2	17.4/8.7	8.6/1.4	18.7/14.4	5.3/3.1	5.0/3.7	3.3/0.9	3.1/0.4	20.6/6.2
distil-large-v2	4.1/1.0	38.0/16.0	69.5/29.6	45.6/64.3	22.1/11.2	16.0/7.2	9.2/1.3	18.9/14.3	6.7/3.2	5.2/1.0	3.6/0.7	3.4/0.5	19.2/5.2
hubert-large-ls960-ft	12.4/1.4	58.5/14.3	50.0/11.0	109.8/100.0	44.4/29.5	21.3/2.4	12.6/1.2	30.2/20.4	23.4/7.1	18.7/3.7	3.6/1.3	2.2/0.1	32.2/12.0
distil-medium.en	4.6/0.6	39.3/17.5	71.3/34.0	45.8/63.9	23.6/8.5	15.6/7.1	9.7/1.4	20.0/14.5	8.5/3.4	7.4/3.7	4.3/1.7	4.2/0.9	21.0/5.7
distil-small.en	4.6/0.6	46.8/19.4	77.0/41.6	54.3/73.3	24.2/8.0	15.4/7.8	11.3/2.3	21.5/18.2	11.7/8.1	9.0/4.7	4.1/0.9	4.0/0.6	21.4/6.4
whisper-medium.en	4.3/1.0	34.2/18.8	66.6/22.8	43.3/60.4	23.0/11.3	19.4/9.5	8.4/1.5	21.3/15.4	4.8/2.3	5.7/3.7	3.5/0.9	3.1/0.4	20.6/5.7
whisper-small.en	4.1/0.9	38.7/19.7	68.8/29.2	50.9/72.9	24.5/13.7	20.8/10.7	9.4/1.2	20.9/16.4	9.6/6.4	7.2/5.8	3.7/0.9	3.6/0.3	21.5/6.7
hf-seamless-m4t-medium	13.2/4.4	57.9/32.1	52.7/50.7	51.4/67.1	57.0/43.7	50.3/24.9	8.8/2.1	36.0/31.3	15.9/16.1	6.4/2.1	8.9/2.3	3.7/0.5	46.1/26.3
whisper-tiny	8.8/3.6	122.1/41.9	110.3/63.9	88.0/90.2	40.3/26.3	22.5/12.3	15.6/5.5	38.6/31.6	54.7/50.6	20.0/13.6	10.8/7.1	7.6/1.7	32.8/16.8
whisper-large	3.7/0.6	48.1/15.8	65.7/18.5	37.2/55.7	22.6/12.9	18.1/9.1	8.5/1.0	18.6/14.8	4.2/2.5	4.0/2.6	3.1/0.7	3.0/0.1	20.0/5.7
whisper-large-v2	4.3/0.8	34.1/17.9	67.1/19.6	38.9/57.6	24.1/15.8	18.2/11.2	8.4/1.1	23.6/15.5	4.4/3.6	3.2/1.0	2.7/0.6	3.0/0.3	20.0/6.9
whisper-large-v3-turbo	3.4/0.4	31.7/14.7	66.2/18.7	34.5/49.8	23.8/10.0	15.7/7.4	7.8/0.9	18.5/13.8	3.9/1.7	4.7/2.6	2.7/0.5	3.3/0.1	20.0/6.0
Qwen2-Audio-7B	4.6/2.7	36.3/15.6	44.8/35.7	31.7/46.7	35.7/14.9	47.4/32.1	5.5/1.3	35.3/37.1	23.3/7.0	5.9/5.8	2.3/1.3	2.0/0.7	25.5/22.8

Table 2: WER and coarse-grained HER across different models and datasets. The values are presented as WER/HER. The lowest HER for each dataset is highlighted in green. Abbreviations: SPGI (SPGISpeech), ATCOSIM (ATCOsim Corpus), ADV (Adversarial), AMI (AMI Corpus), SLU (SLUE-VoxCeleb), SC (Supreme-Court-Speech), SALT (SALT Multispeaker English), LS_Noise (LibriSpeech Test Noise), LS (LibriSpeech ASR Test).

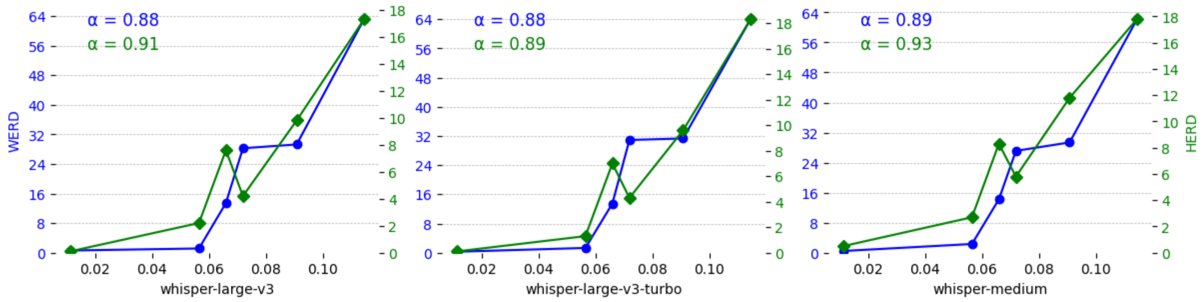


Figure 2: Degradation in word error rate WERD (blue, left y-axis) and hallucinated error rate HERD (green, right y-axis) w.r.t distribution shift (x-axis), measured using Central Moment Discrepancy (CMD) for three different models. The correlation factor, α, is represented by the color, which corresponds to the type of error. Each point on the line represents a new target domain.

Evaluation Pair	Agreement Score (±)
Human & Human 2	0.71 ± 0.01
Human & heuristic	0.00 ± 0.00
Human 2 & GPT_coarsegrained	0.60 ± 0.01
Human 2 & Gemini-2.0	0.59 ± 0.01
GPT_coarsegrained & heuristic	0.10 ± 0.00
GPT_coarsegrained & Gemini-2.0	0.78 ± 0.01
heuristic & Gemini-2.0	0.14 ± 0.01

Table 3: Raw agreement scores between different hallucination evaluation methods.

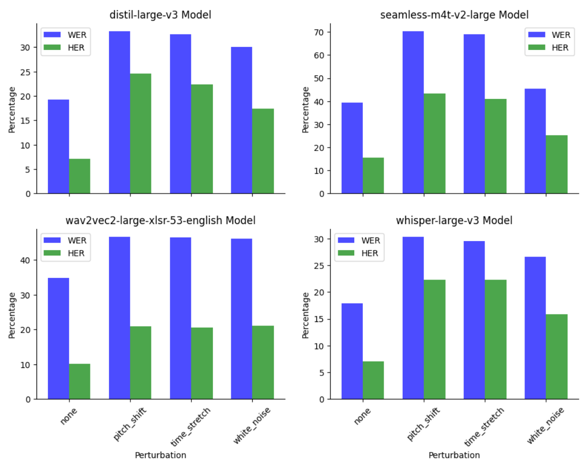


Figure 3: HER and WER for various perturbations across four models.

time stretching lead to a substantial increase of approximately (242%) in both WER and HER across these models, while white noise causes smaller degradations of approximately (142%). Interestingly, the supervised *wav2vec2* model exhibits non-uniform behavior, where the impact on WER and HER is similar across all perturbations. Furthermore, it is noteworthy that HER increased by 50% in the *wav2vec2* model, which is considerably less than the sharp increase observed in *whisper* and *seamless* models, highlighting a notable contrast in the robustness of these models to random synthetic shifts against more targeted shifts.

5.3 Impact of Model Architecture and Scale on Error Rates

Model Type. We hypothesize that model architecture plays a critical role in influencing error rates, with these effects further modulated by the scale and diversity of the training data. As shown in Appendix Table 8, our findings indicate that encoder-only models *Wav2vec2-large-xlsr-53-english* and *Hubert-large-ls960-ft* exhibit the lowest HER/WER ratios across multiple datasets, particularly when compared to models like *Qwen2-Audio-7B* and *hf-seamless-m4t-large*. This suggests that encoder-only models are less prone to hallucinations relative to their overall errors, likely due to robust training or architectural advantages.

Model Size. In our experiments, we evaluate models of varying sizes where training data is fixed across different sizes (ie: *whisper*). More specifically, we select 14 models from the *whisper* family, ranging from the smallest *whisper-tiny* (39M parameters) to the largest *whisper-large-v3* (1.5B parameters). We then calculate both WER and HER for each of these models, following the methodology outlined in Section 3. Our findings show that for smaller models, such as *whisper-tiny* and *whisper-small*, there is a significant increase in both WER and HER. While larger models such as *whisper-medium* and *whisper-large* show substantial improvements, as shown in Figure 4. However, this reduction is not linear. After a certain point, performance improvements in both metrics become less pronounced. This non-monotonic behavior is particularly evident when comparing models in the mid-range of parameter sizes, such as *whisper-medium* and *whisper-large-v3-turbo*, where the difference in performance becomes marginal despite the difference in model size. In conclusion, while larger models generally result in lower WER and HER, the benefits of scaling up model size diminish beyond a certain point, at least when it comes to more nuanced error types.

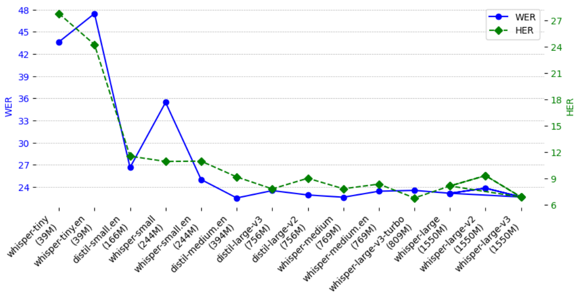


Figure 4: HER (green, right y-axis) and WER (blue, left y-axis) averaged across all datasets. X-axis denotes models with different size in increasing order.

6 Conclusion

In our work, we introduce the Hallucination Error Rate (HER) as a crucial complement to traditional ASR evaluation metrics like WER, especially in high-risk applications where model reliability is critical. By developing a robust LLM-based hallucination detection framework, we present a comprehensive evaluation of ASR models across both synthetic and natural distribution shifts, highlighting the specific challenges ASR systems face under real-world conditions. Our findings emphasize the importance of incorporating HER into standard ASR evaluation practices, particularly for applications in safety-critical domains such as healthcare, legal, and aviation. Through detailed analysis, we show that traditional metrics like WER can mask significant hallucinations, emphasizing the need for more holistic evaluation methods. Our work lays the ground for future advancements in ASR model reliability, aiming to ensure that ASR systems not only produce accurate transcriptions but also avoid generating misleading, harmful, and unfaithful to input speech transcriptions. In future work, we plan to expand our evaluation to cover additional evaluation setups, ensuring a comprehensive set of assessments. Additionally, we aim to explore mitigation strategies, which are critical for enhancing the reliability of ASR systems.

7 Limitations

In this work, we explore the hallucination phenomenon in ASR systems, particularly focusing on the potential causes such as distribution shift, model types, and model size model architectures. While our work offers valuable insights into model behavior across different conditions, there are several limitations to consider.

Evaluation Datasets. In our study, ASR models were evaluated across multiple domains, including legal, medical, and conversational speech, ensuring a broad range of datasets not seen during training. However, it is possible that some of the datasets we treat as target domains may have been inadvertently exposed to the models during training. Furthermore, the lack of access to a diverse variety of domain-specific datasets limits our understanding of how these models will perform in more diverse or previously unseen domains, particularly those with limited or noisy data. While our focus on domain shifts is an important step, further research is needed to assess model performance in even more varied and challenging real-world environments.

Synthetic Noise and Perturbations. Our experiments also include synthetic noise and perturbations to evaluate model robustness. While this approach helps simulate real-world challenges, it does not capture all possible distortions that may occur in uncontrolled environments. Adversarial noise, pitch shifts, and time stretching are some of the perturbations we consider, but other potential real-world disruptions, such as cross-lingual noise or complex acoustic reverberations, are not fully explored.

Hallucination Detection. The detection of hallucinations in ASR systems, as measured by the Hallucination Error Rate (HER), is a key contribution of our study. However, our reliance on LLM-based classifiers introduces potential biases and variability. While we observe strong alignment with human judgments, the accuracy of these evaluations may be influenced by subjective interpretation, especially in edge cases where the boundaries between errors are unclear. Additionally, while LLM-based methods present a novel approach, their performance in low-resource settings or with models trained on smaller datasets has not been fully explored. Furthermore, the use of proprietary models, such as those from OpenAI via API, introduces additional costs, which could limit the scalability of this approach.

8 Ethics Statement

Data Collection and Release. For this study, we rely on publicly available datasets from diverse domains to evaluate hallucinations in ASR systems. We ensure that the data used in our research is appropriately sourced, maintaining respect for copyright, license, and privacy regulations. Furthermore, we emphasize that the use of these datasets is strictly for academic purposes, aligned with the principles of fair use.

Intended Use. Our work aims to enhance the robustness of ASR systems, especially in high-stakes environments where errors can have significant consequences. We believe our findings will encourage further research in hallucination detection, with particular attention to models' performance in low-resource and critical domains such as healthcare and law. By introducing the Hallucination Error Rate (HER) as a complementary metric to traditional evaluation methods, we hope to inspire the development of more reliable and transparent ASR systems.

Potential Misuse and Bias. While our work provides valuable insights about hallucinations in ASR systems, we acknowledge that they could be misused if deployed in inappropriate contexts. Since these models are trained on a variety of data sources, there is the potential for them to generate biased or harmful content, especially if the training data contains any inherent biases. Moreover, hallucinations in ASR outputs, if undetected, can lead to severe consequences in critical applications such as legal, medical, and financial settings. We recommend careful deployment of these models, ensuring that they undergo rigorous bias mitigation and hallucination detection processes before being used in such domains.

References

Ayo Adedeji, Sarita Joshi, and Brendan Doohan. 2024. The sound of healthcare: Improving medical transcription asr accuracy with large language models. *arXiv preprint arXiv:2402.07658*.

Tejumade Afonja, Tobi Olatunji, Sewade Ogun, Naome A. Etori, Abraham Owodunni, and Moshood Yekini. 2024. Performant asr models for medical entities in accented speech. *Preprint*, arXiv:2406.12387.

DSMTF American Psychiatric Association, DS American Psychiatric Association, et al. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5*, volume 5. American psychiatric association Washington, DC.

Carlos Arriaga, Alejandro Pozo, Javier Conde, and Alvaro Alonso. 2024. Evaluation of real-time transcriptions using end-to-end asr models. *arXiv preprint arXiv:2409.05674*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *Preprint*, arXiv:2404.18930.

Mateusz Barański, Jan Jasiński, Julitta Bartolewska, Stanisław Kacprzak, Marcin Witkowski, and Konrad Kowalczyk. 2025. Investigation of whisper asr hallucinations induced by non-speech audio. *arXiv preprint arXiv:2501.11378*.

Mateusz Barański, Jan Jasiński, Julitta Bartolewska, Stanisław Kacprzak, Marcin Witkowski, and Konrad Kowalczyk. 2025. Investigation of whisper asr hallucinations induced by non-speech audio. *Preprint*, arXiv:2501.11378.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Seamless Communication, Loïc Barrault, Yu-Anh Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. Seamlessm4t: Massively multilingual multimodal machine translation. *Preprint*, arXiv:2308.11596.

The AMI Consortium. n.d. Ami corpus. Accessed: 2025-02-04.

Miguel Del Rio, Natalie Delworth, Ryan Westerman, Michelle Huang, Nishchal Bhandari, Joseph Palakapilly, Quinten McNamara, Joshua Dong, Piotr Żelasko, and Miguel Jetté. 2021. Earnings-21: A practical benchmark for asr in the wild. In *Interspeech 2021*, interspeech$_2021, page3465˘3469. ISCA$.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: sentence-level multimodal and language-agnostic representations. *arXiv e-prints*, pages arXiv–2308.

Qi Fan, Mattia Segu, Yu-Wing Tai, Fisher Yu, Chi-Keung Tang, Bernt Schiele, and Dengxin Dai. 2022. Normalization perturbation: A simple domain generalization method for real-world domain shifts. *arXiv preprint arXiv:2211.04393*.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. 2024. Do i know this entity? knowledge awareness and hallucinations in language models. *arXiv preprint arXiv:2411.14257*.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. *arXiv preprint arXiv:2010.05873*.

Rita Frieske and Bertram E Shi. 2024. Hallucinations in neural automatic speech recognition: Identifying errors and hallucinatory models. *arXiv preprint arXiv:2401.01572*.

Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush. 2023. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *Preprint*, arXiv:2311.00430.

Nicolad Garneau and Olivier Bolduc. 2024. The state of commercial automatic french legal speech recognition systems and their impact on court reporters et al. *Preprint*, arXiv:2408.11940.

Stephanie M Hare. 2021. Hallucinations: A functional network model of how sensory representations become selected for conscious awareness in schizophrenia. *Frontiers in Neuroscience*, 15:733038.

Konrad Hofbauer, Stefan Petrik, and Horst Hering. 2008. The ATCOSIM corpus of non-prompted clean air traffic control speech. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Jaeyoung Huh, Sangjoon Park, Jeong Eun Lee, and Jong Chul Ye. 2023. Improving medical speech-to-text accuracy with vision-language pre-training model. *Preprint*, arXiv:2303.00091.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on large language model hallucination via a creativity perspective. *Preprint*, arXiv:2402.06647.

Adam Tauman Kalai and Santosh S Vempala. 2024. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 160–171.

Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, and Roger Zimmermann. 2020. Domain divergences: A survey and empirical analysis. *arXiv preprint arXiv:2010.12198*.

Seung-Eun Kim, Bronya R Chernyak, Olga Seleznova, Joseph Keshet, Matthew Goldrick, and Ann R Bradlow. 2024. Automatic recognition of second language speech-in-noise. *JASA Express Letters*, 4(2).

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. 2024a. Careless whisper: Speech-to-text hallucination harms. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1672–1681. ACM.

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X Mei, Hilke Schellmann, and Mona Sloane. 2024b. Careless whisper: Speech-to-text hallucination harms. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1672–1681.

Feng-Ting Liao, Yung-Chieh Chan, Yi-Chang Chen, Chan-Jan Hsu, and Da-Shan Shiu. 2023. Zero-shot domain-sensitive speech recognition with prompt-conditioning fine-tuning. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2021. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Ashish Mittal, Rudra Murthy, Vishwajeet Kumar, and Riyaz Bhat. 2024. Towards understanding and mitigating the hallucinations in nlp and speech. In *Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD)*, pages 489–492.

Dena Mujtaba, Nihar R Mahapatra, Megan Arney, J Scott Yaruss, Hope Gerlach-Houck, Caryn Herring, and Jia Bin. 2024. Lost in transcription: Identifying and quantifying the accuracy biases of automatic speech recognition systems against disfluent speech. *arXiv preprint arXiv:2405.06150*.

Raphael Olivier and Bhiksha Raj. 2022. Recent improvements of asr models in the face of adversarial attacks. *Interspeech*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. `http://www.openslr.org/12/`. Accessed: [Insert Date].

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. PriMock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598, Dublin, Ireland. Association for Computational Linguistics.

Ankur P Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. *arXiv preprint arXiv:2004.14373*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Shangeth Rajaa and Abhinav Tushar. SpeechLLM: Multi-Modal LLM for Speech Understanding.

Hadeel Saadany, Catherine Breslin, Constantin Orăsan, and Sophie Walker. 2022. Better transcription of uk supreme court hearings. *Preprint*, arXiv:2211.17094.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A comprehensive survey of hallucination in large language, image, video and audio foundation models. *Preprint*, arXiv:2405.09589.

Zitha Sasindran, Harsha Yelchuri, and TV Prabhakar. 2024. Semascore: a new evaluation metric for automatic speech recognition tasks. *arXiv preprint arXiv:2401.07506*.

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

Prashant Serai, Vishal Sunder, and Eric Fosler-Lussier. 2021. Hallucination of speech recognition errors with sequence to sequence learning. *Preprint*, arXiv:2103.12258.

Prashant Serai, Vishal Sunder, and Eric Fosler-Lussier. 2022. Hallucination of speech recognition errors with sequence to sequence learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:890–900.

Syed Abdul Gaffar Shakhadri, Kruthika KR, and Kartik Basavaraj Angadi. 2025. Samba-asr state-of-the-art speech recognition leveraging structured state-space models. *arXiv preprint arXiv:2501.02832*.

Suwon Shon, Ankita Pasad, Felix Wu, Pablo Brusco, Yoav Artzi, Karen Livescu, and Kyu J Han. 2022. Slue: New benchmark tasks for spoken language understanding evaluation on natural speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7927–7931. IEEE.

Matthias Sperber, Hendra Setiawan, Christian Gollan, Udhyakumar Nallasamy, and Matthias Paulik. 2020. Consistent transcription and translation of speech. *Transactions of the Association for Computational Linguistics*, 8:695–709.

Adam Stiff, Prashant Serai, and Eric Fosler-Lussier. 2019. Improving human-computer interaction in low-resource settings with text-to-phonetic data augmentation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7320–7324. IEEE.

Vishal Sunder, Prashant Serai, and Eric Fosler-Lussier. 2022. Building an asr error robust spoken virtual patient system in a highly class-imbalanced scenario without speech data. *Preprint*, arXiv:2204.05183.

Piotr Szymański, Lukasz Augustyniak, Mikolaj Morzy, Adrian Szymczak, Krzysztof Surdyk, and Piotr Żelasko. 2023. Why aren't we NER yet? artifacts of ASR errors in named entity recognition in spontaneous speech transcripts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1746–1761, Toronto, Canada. Association for Computational Linguistics.

Kensho Technologies. 2021. Spgispeech: A large-scale, high-quality dataset for speech recognition in financial earnings calls. `https://datasets.kensho.com/datasets/spgispeech`. Accessed: [Insert Date].

Prathiksha Rumale Vishwanath, Simran Tiwari, Tejas Ganesh Naik, Sahil Gupta, Dung Ngoc Thai, Wenlong Zhao, SUNJAE KWON, Victor Ardulov, Karim Tarabishy, Andrew McCallum, and Wael Salloum. 2024. Faithfulness hallucination detection in healthcare AI. In *Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare*.

Wenbin Wang, Yang Song, and Sanjay Jha. 2024. Globe: A high-quality english corpus with global accents for zero-shot speaker adaptive text-to-speech. *arXiv preprint arXiv:2406.14875*.

Steven M Williamson and Victor Prybutok. 2024. The era of artificial intelligence deception: Unraveling the complexities of false realities and emerging threats of misinformation. *Information*, 15(6):299.

Qianqian Xie, Edward Schenck, He Yang, Yong Chen, Yifan Peng, and Fei Wang. 2023. Faithful ai in medicine: A systematic review with large language models and beyond.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.

Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. 2019. Central moment discrepancy (cmd) for domain-invariant representation learning. *Preprint*, arXiv:1702.08811.

Jun Zhang, Jingyue Wu, Yiyi Qiu, Aiguo Song, Weifeng Li, Xin Li, and Yecheng Liu. 2023. Intelligent speech technologies for transcription, disease diagnosis, and medical equipment interactive control in smart hospitals: A review. *Computers in Biology and Medicine*, 153:106517.

Leor Zmigrod, Jane R Garrison, Joseph Carr, and Jon S Simons. 2016. The neural mechanisms of hallucinations: a quantitative meta-analysis of neuroimaging studies. *Neuroscience & Biobehavioral Reviews*, 69:113–123.

A Appendix

A.1 Experiments

A.1.1 Datasets

A.1.2 Perturbation

To evaluate the robustness of ASR models under varying conditions, we apply the following perturbations to the audio inputs:

- *White Noise*: Gaussian noise is added at a low amplitude to simulate environmental interference.
- *Time Stretching*: The audio is randomly stretched by a factor between 0.9 and 1.1, altering the speed while preserving pitch.
- *Pitch Shifting*: The pitch is randomly shifted by up to ±2 semitones to mimic natural variations in speech.
- *None*: No perturbation is applied, serving as the baseline for comparison.

These perturbations are designed to replicate real-world challenges such as background noise, speaker variability, and recording inconsistencies.
.

A.1.3 Models

A.2 Prompts

A.3 Results

We provide a detailed analysis of the experimental results, focusing on the highest and lowest performing models across the study. To offer a comprehensive overview, we present Hallucination Error Rate (HER) and Word Error Rate (WER) across domain shifts for all models, as shown in . Additionally, we include fine-grained error analysis which highlights the differences between coarse-grained and fine-grained error categorization.

We also calculate HER to WER ratio. As robust models would exhibit a smaller gap between hallucination and non-hallucination errors.

Furthermore, we present the percentage of non-hallucination errors across datasets and models, categorizing them into Phonetic (P), Oscillation (O), and Language (L) errors. This analysis provides deeper insights into the types of errors that are most frequent and their distribution across different dataset-model combinations.

We also highlight the overall distribution across all datasets and the robustness of both levels (coarsegrained and finegrained) in correctly identifying hallucination.

Key Findings:

- The highest and lowest performing models exhibit significant variations in HER and WER under domain shifts, with some models showing robustness while others struggle.
- Fine-grained error analysis reveals that certain error types (e.g., Oscillation) are more prevalent in specific dataset-model combinations.
- Non-hallucination errors, particularly Phonetic and Language errors, dominate in certain scenarios, providing actionable insights for improving model performance.

These results underscore the importance of considering both hallucination and non-hallucination errors when evaluating ASR systems, as well as the need for domain-specific adaptations to enhance robustness.

Name	Domain	Recording Conditions	Description
LibriSpeech	Speech Recognition	High-quality, read speech from audiobooks	A corpus of approximately 1,000 hours of 16kHz read English speech, derived from LibriVox audiobooks, segmented and aligned for ASR tasks.
GLOBE	Accented Speech	Close-talk microphone	Contains utterances from 23,519 speakers and covers 164 accents worldwide, recorded in close-talk microphone conditions.
Supreme-court	Legal	Diverse acoustic conditions (audiobooks, podcasts, YouTube)	A multi-domain, multi-style speech recognition corpus incorporating diverse acoustic and linguistic conditions, sourced from audiobooks, podcasts, and YouTube.
SPGISpeech	Finance	Corporate earnings calls (professional transcription)	Contains 5,000 hours of professionally transcribed audio from corporate earnings calls, featuring both spontaneous and narrated speaking styles.
Adversarial	Synthetic	Corporate earnings calls	Includes multiple splits with utterances modified using adversarial noise of varying radii (0.04 and 0.015) and combined with Room Impulse Response (RIR) noise.
AMI (IHM)	Meetings	Multi-device meeting environment	The AMI Meeting Corpus is a 100-hour dataset of English meeting recordings, featuring multimodal data synchronized across various devices.
SLUE - VoxCeleb	Conversational	YouTube video extracts (conversational)	Consists of single-sided conversational voice snippets extracted from YouTube videos, originally designed for speaker recognition.
Primock57	Medical	Mock consultations by clinicians	Contains mock consultations conducted by seven clinicians and 57 actors posing as patients, representing a diverse range of ethnicities, accents, and ages.
BERSt	Home Environment	Home recordings using smartphones	A collection of speech data recorded in home environments using various smartphone microphones, with participants from diverse regions.
ATCOsim	Aviation	Real-time air traffic control simulations	A specialized database containing ten hours of English speech from ten non-native speakers, recorded during real-time air traffic control simulations.

Table 4: Speech Datasets for ASR, categorized by domain, recording conditions, and description.

Type	Description
None	We do not apply any modification to the input speech. Serving as the baseline for comparison.
White Noise	Gaussian noise is added at a low amplitude, simulating environmental interference.
Time Stretching	The audio is randomly stretched by a factor between 0.9 and 1.1, altering the speed without affecting the pitch.
Pitch Shifting	The pitch is randomly shifted by up to ±2 semitones.

Table 5: Perturbation details for synthetic shift.

Model Type and Models	Parameters	Architecture	Pre-Training Objective and Training Data
wav2vec2 (Baevski et al., 2020) – wav2vec2-large-xlsr-53-english	315M	7-Conv (Kernel 10/3/3/3/3/2/2) + 24-Trans	Self-supervised pre-training on raw audio via contrastive loss. Training data: Common Voice 6.1 (53 languages).
hubert (Hsu et al., 2021) – hubert-large-ls960-ft	316M	7-Conv (Kernel 10/3/3/3/3/2/2) + 24-Trans	Masked prediction pre-training. Training data: Libri-Light (60k hours).
seamless (Seamless Communication, 2023) – hf-seamless-m4t-large – hf-seamless-m4t-medium	2.3B 1.2B	UnitY2 (Enc-Dec + Text Decoder)	Multilingual ASR/translation. Training data: 443k hours of aligned speech-text (29 languages).
speechllm (Rajaa and Tushar) – speechllm-1.5B	1.5B	HubertX encoder + TinyLlama decoder	Audio-text alignment via multi-task learning. Training data: Proprietary ASR datasets.
whisper (Radford et al., 2022) – whisper-large-v3 – distil-large-v3 – whisper-large-v2 – whisper-large-v3-turbo – distil-large-v2 – whisper-large – whisper-tiny – whisper-tiny.en – whisper-medium – whisper-medium.en – distil-medium.en – distil-small.en – whisper-small – whisper-small.en	1.55B 756M 769M 244M 39M	2-Conv (Kernel 3x3, stride 2) + 32-Trans (large) 2-Conv + 24-Trans (medium) 2-Conv + 12-Trans (small)	Multilingual ASR/translation. Pre-training: 680k hours of web-crawled audio.
Qwen (Chu et al., 2024) – Qwen2-Audio-7B	7B	Audio encoder + QwenLM decoder	Multi-task pretraining (ASR, TTS, alignment). Training data: 3M audio-text pairs.

Table 6: Model architectures, parameters, and training details. Whisper variants include convolutional layers for spectrogram downsampling.

You are a classifier trained to detect and categorize specific transcription errors produced by a speech recognition system. The possible categories are:
1. **Hallucination Error**: The output contains fabricated, contradictory, or invented information that is not supported by the ground truth. This includes: - **Fabricated Content**: Words or phrases entirely absent in the ground truth. - **Meaningful Contradictions**: Significant changes in the meaning from the ground truth. - **Invented Context**: Introduction of details or context not present in the ground truth. - **Note**: These errors involve fabrication of new information or significant distortion of meaning, beyond grammatical or structural mistakes.
2. **Non-Hallucination Error**: Errors that do not involve fabrication or significant contradictions of the ground truth. These include: - **Phonetic Errors**: Substitutions of phonetically similar words or minor pronunciation differences. - **Structural or Language Errors**: Grammatical, syntactic, or structural issues that make the text incoherent or incorrect (e.g., incorrect verb tenses, subject-verb agreement problems, omissions, or insertions). - **Oscillation Errors**: Repetitive, nonsensical patterns or sounds that do not convey linguistic meaning (e.g., "ay ay ay ay"). - **Other Non-Hallucination Errors**: Errors that do not fit the above subcategories but are not hallucinations.
3. **No Error**: The generated output conveys the same meaning as the ground truth, even if the phrasing, grammar, or structure differs. Minor differences in wording, phrasing, or grammar that do not alter the intended meaning are acceptable.

Input Format:
Ground Truth: The original, accurate text provided.
Generated Output: The text produced by the speech recognition system.

Output Format: Classify the input text pairs into one of the following:
Non-Hallucination Error
Hallucination Error
No Error

Examples:
Example 1:
Ground Truth: "A millimeter roughly equals one twenty-fifth of an inch."
Generated Output: "Miller made her roughly one twenty-fifths of an inch."
Output: Non-Hallucination Error

Example 2:
Ground Truth: "Indeed, ah!"
Generated Output: "Ay ay indeed ay ay ay ay ay ay."
Output: Non-Hallucination Error

Example 3:
Ground Truth: "Captain Lake did not look at all like a London dandy now."
Generated Output: "Will you let Annabel ask her if she sees what it is you hold in your arms again?"
Output: Hallucination Error

Example 4:
Ground Truth: "The patient was advised to take paracetamol for fever and rest for two days."
Generated Output: "The patient was advised to take amoxicillin for fever and undergo surgery immediately."
Output: Hallucination Error

Example 5:
Ground Truth: "I need to book a flight to New York."
Generated Output: "I need to book ticket to New York."
Output: No Error

Example 6:
Ground Truth: "She went to the store yesterday."
Generated Output: "She went to the shop yesterday."
Output: No Error

Instruction: You must produce only the classification as the output. Do not include explanations, reasoning, or additional information.

Input: Ground Truth: "{ground_truth}" Generated Output: "{output}"
Output: {{insert your classification here}}

Figure 5: Coarsegrained error detection prompt. The task is to classify transcription errors produced by an ASR model into one of three categories: *Non-Hallucination Error*, *Hallucination Error*, or *No Error*.

You are a classifier trained to detect and categorize specific transcription errors produced by a speech recognition system. The possible categories are:
1. **Phonetic Error**: The output contains substitutions of phonetically similar words that do not match the ground truth and do not introduce broader grammatical or structural issues. These errors typically involve misrecognition of similar-sounding words or minor pronunciation differences.
2. **Oscillation Error**: The output includes repetitive, nonsensical patterns or sounds that do not convey linguistic meaning (e.g., "ay ay ay ay").
3. **Hallucination Error**: The output contains fabricated, contradictory, or invented information that is not supported by the ground truth. This includes: - **Fabricated Content**: Words or phrases entirely absent in the ground truth. - **Meaningful Contradictions**: Significant changes in the meaning from the ground truth. - **Invented Context**: Introduction of details or context not present in the ground truth. - **Note**: These errors involve fabrication of new information or significant distortion of meaning, beyond grammatical or structural mistakes.
4. **Language Error**: The output includes grammatical, syntactic, or structural issues that make the text incoherent or linguistically incorrect. This category encompasses errors such as: - Incorrect verb tenses or subject-verb agreement problems. - Sentence fragments or incomplete structures. - Omissions or insertions of words that do not fabricate new context. - Incomplete sentences or phrases that do not convey the intended meaning as ground truth. - **Note**: Incomplete sentences or phrases are classified as Language Errors only when they do not fabricate new meaning or deviate from the intent of the ground truth.
5. **No Error**: The generated output conveys the same meaning as the ground truth, even if the phrasing, grammar, or structure differs. Minor differences in wording, phrasing, punctuation, or casing that do not alter the intended meaning are not considered errors. - **Note**: Minor omissions, such as missing articles, are acceptable as long as they do not change the meaning of the ground truth.

Input Format:
Ground Truth: The original, accurate text provided.
Generated Output: The text produced by the speech recognition system.

Output Format: Classify the input text pairs into one of the following:
Phonetic Error
Oscillation Error
Hallucination Error
Language Error
No Error

Examples:
Example 1:
Ground Truth: "A millimeter roughly equals one twenty-fifth of an inch."
Generated Output: "Miller made her roughly one twenty-fifths of an inch."
Output: Phonetic Error

Example 2:
Ground Truth: "I will go to New York City!"
Generated Output: "Ay ay ay ay ay ay ay ay."
Output: Oscillation Error

Example 3:
Ground Truth: "Captain Lake did not look at all like a London dandy now."
Generated Output: "Will you let Annabel ask her if she sees what it is you hold in your arms again?"
Output: Hallucination Error

Example 4:
Ground Truth: "The cat is chasing the mouse."
Generated Output: "The cat chased by the mouse."
Output: Language Error

Example 5:
Ground Truth: "I need to book a flight to New York."
Generated Output: "I need to book ticket to New York."
Output: No Error

Your Task: Classify the input into one of the five categories.
Instruction: You must produce only the classification as the output. Do not include explanations, reasoning, or additional information.

Input: Ground Truth: "{ground_truth}" Generated Output: "{output}"
Output: {{insert your classification here}}

Figure 6: Finegrained error detection prompt. The task is to classify transcription errors produced by an ASR model into one of five categories: *Phonetic Error*, *Oscillation Error*, *Hallucination Error*, *Language Error*, or *No Error*.

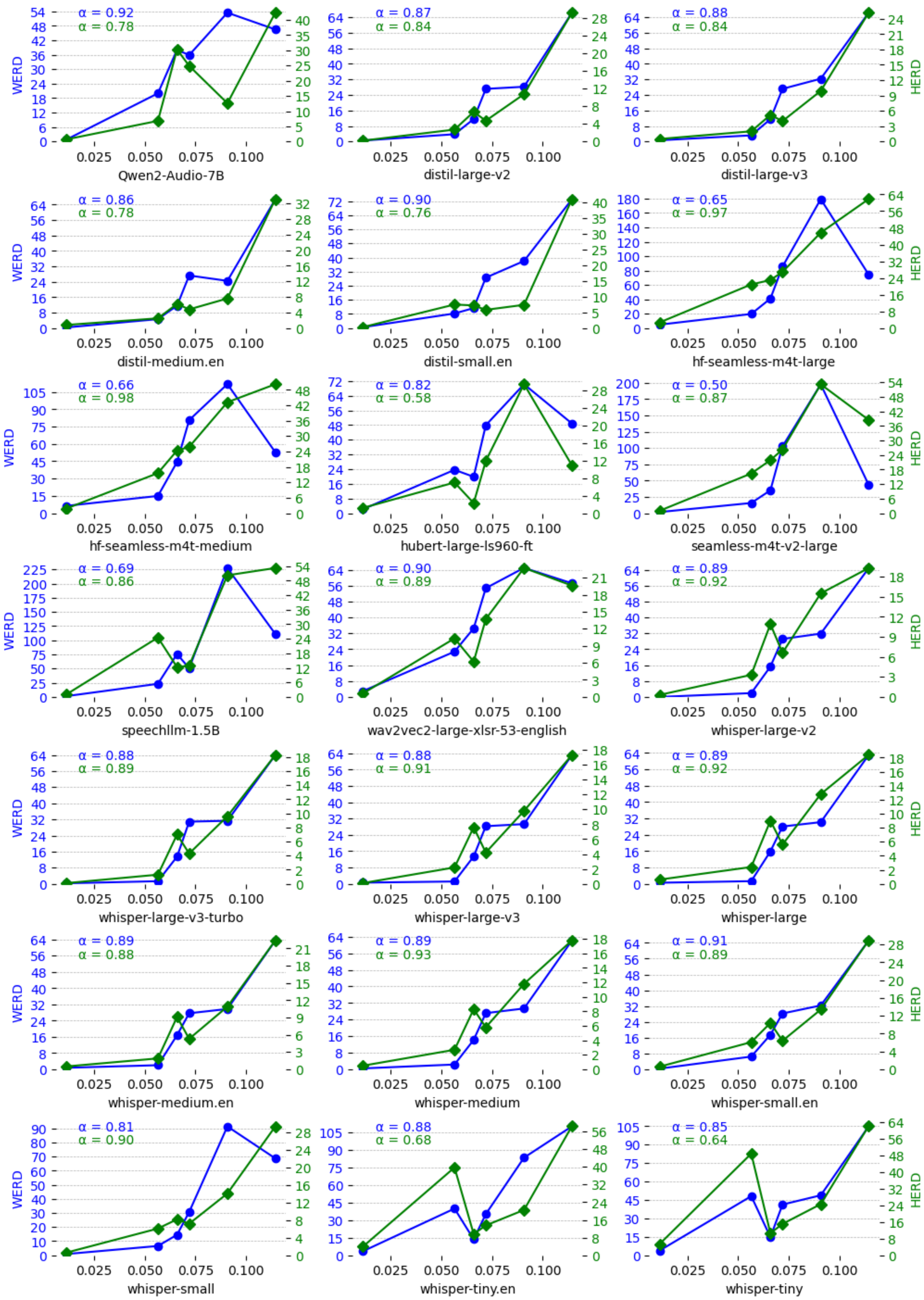


Figure 7: Degradation in WER (blue, left y-axis) and HER (green, right y-axis) w.r.t distribution shift (x-axis), measured using Central Moment Discrepancy (CMD) for all models.

Model	SPGI	BERSt	ATCOsim	ADV	AMI	SLU	SNIPS	SC	GLOBE	SALT	LS_Noise	LS	Primock57
whisper-large-v3	3.4/0.9	32.4/13.2	65.3/13.1	33.3/47.1	23.4/11.3	15.5/13.4	8.2/0.5	18.8/14.8	3.4/1.9	3.0/1.0	2.6/0.4	2.2/0.5	19.2/4.8
wav2vec2-large-xlsr-53-english	19.6/2.9	64.0/13.9	63.0/11.9	100.1/95.7	53.0/22.8	43.4/14.6	12.4/0.9	32.6/17.9	27.0/9.3	17.0/2.1	9.0/0.8	6.5/0.1	47.9/15.8
hf-seamless-m4t-large	14.7/4.5	58.9/29.5	76.6/55.1	61.2/71.4	63.7/43.2	44.0/27.5	7.4/2.6	34.2/30.8	19.2/19.2	4.2/1.0	6.5/2.9	3.4/0.3	44.5/25.8
speechllm-1.5B	11.5/4.1	68.5/31.8	121.4/38.3	95.3/92.5	127.3/52.3	83.8/19.3	10.8/2.8	41.3/36.5	27.9/21.8	9.5/4.2	10.9/4.3	11.4/4.2	41.7/17.3
whisper-medium	3.7/1.1	34.5/14.5	65.6/14.4	42.9/58.4	23.2/11.4	17.4/13.6	8.6/1.1	18.7/14.0	5.3/2.9	5.0/3.7	3.3/0.8	3.1/0.2	20.6/5.9
distil-large-v2	4.1/0.9	38.0/14.7	69.5/22.4	45.6/60.4	22.1/13.5	16.0/13.4	9.2/0.8	18.9/15.0	6.7/3.0	5.2/1.0	3.6/0.5	3.4/0.3	19.2/5.9
hubert-large-ls960-ft	12.4/2.0	58.5/11.3	50.0/6.4	109.8/100.0	44.4/28.1	21.3/6.5	12.6/1.3	30.2/23.8	23.4/6.5	18.7/4.2	3.6/2.1	2.2/0.1	32.2/11.7
distil-medium.en	4.6/0.6	39.3/14.8	71.3/26.8	45.8/58.8	23.6/10.6	15.6/11.3	9.7/1.1	20.0/15.2	8.5/1.9	7.4/3.1	4.3/0.9	4.2/0.5	21.0/5.4
distil-small.en	4.6/1.0	46.8/17.9	77.0/32.7	54.3/68.6	24.2/10.9	15.4/13.1	11.3/1.7	21.5/18.5	11.7/6.4	9.0/4.7	4.1/0.8	4.0/0.4	21.4/6.8
whisper-medium.en	4.3/1.5	34.2/15.2	66.6/16.2	43.3/58.0	23.0/11.2	19.4/16.3	8.4/1.4	21.3/16.3	4.8/1.7	5.7/4.2	3.5/0.7	3.1/0.4	20.6/6.0
whisper-small.en	4.1/1.3	38.7/17.5	68.8/19.3	50.9/69.8	24.5/13.5	20.8/15.9	9.4/1.1	20.9/16.5	9.6/4.8	7.2/4.7	3.7/0.9	3.6/0.4	21.5/6.7
speecht5$_asr$	25.8/28.1	108.1/32.3	81.7/57.4	117.2/100.0	462.5/25.1	129.4/21.9	24.6/7.8	156.7/43.9	60.1/54.1	53.9/28.3	13.9/14.1	6.0/0.8	53.9/41.2
hf-seamless-m4t-medium	13.2/5.1	57.9/29.1	52.7/40.9	51.4/63.5	57.0/41.9	50.3/25.3	8.8/1.6	36.0/32.6	15.9/14.2	6.4/2.1	8.9/3.0	3.7/0.4	46.1/24.5
whisper-tiny	8.8/4.3	122.1/37.2	110.3/60.3	88.0/85.9	40.3/25.3	22.5/15.4	15.6/4.7	38.6/28.3	54.7/47.6	20.0/13.1	10.8/6.7	7.6/1.6	32.8/15.8
whisper-large	3.7/1.1	48.1/12.8	65.7/14.2	37.2/51.4	22.6/12.8	18.1/15.5	8.5/0.9	18.6/14.5	4.2/1.8	4.0/2.6	3.1/0.5	3.0/0.2	20.0/6.0
whisper-large-v2	4.3/0.9	34.1/14.5	67.1/15.1	38.9/54.1	24.1/14.3	18.2/15.9	8.4/0.5	23.6/15.6	4.4/3.0	3.2/1.6	2.7/0.6	3.0/0.2	20.0/6.4
whisper-large-v3-turbo	3.4/0.9	31.7/11.7	66.2/13.5	34.5/47.1	23.8/10.6	15.7/14.4	7.8/0.6	18.5/14.4	3.9/1.1	4.7/1.6	2.7/0.3	2.5/0.5	20.8/4.8
whisper-tiny.en	6.9/3.0	75.4/32.5	112.3/57.9	80.1/82.7	38.4/20.8	19.4/15.1	14.0/4.1	38.0/28.0	42.1/37.9	19.4/12.6	9.4/5.1	6.1/0.8	31.3/14.8
distil-large-v3	3.7/0.7	33.7/12.0	69.2/18.0	38.6/51.0	23.4/11.7	14.1/12.9	8.8/0.9	19.5/16.7	5.6/1.3	5.0/2.1	3.3/0.6	2.8/0.3	19.1/5.6
whisper-small	4.3/1.2	42.1/17.7	73.4/23.6	77.0/72.5	39.8/14.7	18.2/14.2	9.6/1.4	23.4/17.6	10.0/4.4	7.1/4.2	4.3/0.4	3.7/0.3	22.1/7.4
Qwen2-Audio-7B	4.6/2.7	36.3/15.6	44.8/35.7	31.7/46.7	35.7/14.9	47.4/32.1	5.5/1.3	35.3/37.1	23.3/7.0	5.9/5.8	2.3/1.3	2.0/0.7	25.5/22.8
seamless-m4t-v2-large	15.9/5.4	55.2/25.8	43.5/31.0	50.5/67.5	75.1/50.6	45.9/21.3	6.0/1.6	34.7/24.8	14.9/14.9	5.3/1.6	3.6/1.5	2.7/0.4	37.6/23.5

Table 7: Character Error Rate (CER) and hallucination error rate (HER) across models and datasets. Values are presented as CER/HER.

Model	SPGI	BERSt	ATCOsim	ADV	AMI	SLU	SNIPS	SC	GLOBE	SALT	LS_Noise	LS	Primock57
whisper-large-v3	0.12	0.49	0.27	1.49	0.43	0.51	0.11	0.82	0.74	0.34	0.16	0.14	0.23
wav2vec2-large-xlsr-53-english	0.05	0.29	0.31	0.96	0.43	0.14	0.10	0.45	0.38	0.12	0.07	0.00	0.29
hf-seamless-m4t-large	0.24	0.59	0.81	1.26	0.73	0.53	0.43	0.88	1.11	0.37	0.49	0.15	0.62
speechllm-1.5B	0.42	0.56	0.47	0.99	0.43	0.20	0.27	0.93	1.03	0.44	0.48	0.37	0.41
whisper-medium	0.05	0.47	0.28	1.47	0.53	0.50	0.16	0.77	0.58	0.74	0.28	0.13	0.30
distil-large-v2	0.25	0.42	0.43	1.41	0.51	0.45	0.14	0.76	0.48	0.20	0.20	0.15	0.27
hubert-large-ls960-ft	0.11	0.24	0.22	0.91	0.66	0.11	0.10	0.68	0.30	0.20	0.37	0.05	0.37
distil-medium.en	0.13	0.44	0.48	1.40	0.36	0.45	0.14	0.73	0.40	0.49	0.40	0.21	0.27
distil-small.en	0.13	0.41	0.54	1.35	0.33	0.51	0.20	0.85	0.69	0.53	0.22	0.15	0.30
whisper-medium.en	0.23	0.55	0.34	1.40	0.49	0.49	0.18	0.72	0.47	0.64	0.26	0.13	0.28
whisper-small.en	0.22	0.51	0.42	1.43	0.56	0.51	0.13	0.78	0.67	0.80	0.24	0.08	0.31
hf-seamless-m4t-medium	0.33	0.56	0.96	1.31	0.77	0.50	0.24	0.87	1.01	0.33	0.26	0.13	0.57
whisper-tiny	0.41	0.34	0.58	1.02	0.65	0.55	0.35	0.82	0.93	0.68	0.66	0.22	0.51
whisper-large	0.16	0.33	0.28	1.50	0.57	0.50	0.12	0.80	0.59	0.65	0.23	0.03	0.28
whisper-large-v2	0.19	0.52	0.29	1.48	0.65	0.62	0.13	0.66	0.81	0.33	0.22	0.10	0.34
whisper-large-v3-turbo	0.12	0.46	0.28	1.44	0.42	0.47	0.12	0.75	0.44	0.55	0.19	0.16	0.23
whisper-tiny.en	0.35	0.45	0.53	1.08	0.56	0.55	0.35	0.75	0.97	0.62	0.53	0.18	0.47
distil-large-v3	0.03	0.41	0.37	1.45	0.44	0.40	0.11	0.78	0.43	0.32	0.27	0.14	0.23
whisper-small	0.14	0.46	0.40	0.98	0.36	0.47	0.14	0.74	0.64	0.67	0.21	0.08	0.34
Qwen2-Audio-7B	0.74	0.52	0.96	1.55	0.38	0.66	0.26	1.18	0.33	1.16	0.67	0.39	1.01
seamless-m4t-v2-large	0.41	0.55	0.90	1.42	0.71	0.49	0.29	0.71	1.13	0.40	0.41	0.19	0.71

Table 8: Comparison of HER/WER ratio across models for all datasets.

Model	BERSt			GLOBE			LibriSpeech			Primock57			Adversarial			AMI			ATCOsim			SALT			SLUE			SPGI			SC		
	P	O	L	P	O	L	P	O	L	P	O	L	P	O	L	P	O	L	P	O	L	P	O	L	P	O	L	P	O	L	P	O	L
Q2A-7B	37.41	0.75	5.64	21.3	3.5	8	8.6	0.2	1.3	10.5	5.5	12.5	19.22	0.39	4.31	10	3.4	12.4	49.5	0.7	3.4	11	1.7	2.3	10.1	0.2	1.7	8.38	2.09	4.71	10.7	15.6	24.1
dw-l-v2	35.53	0.19	5.08	22.7	0	4.7	17.7	0.1	5.4	10.2	0.7	14.8	16.08	0	13.73	7.7	0	13.7	57.9	0	2.8	12.5	5.7	4.5	15.9	0	4.2	15.18	0	3.66	11.8	18.5	27.2
dw-l-v3	33.27	0	3.57	21.2	0	3.9	14.3	0.1	3.5	9.3	0.6	13.6	24.71	0	10.59	6.6	0	12.3	61.2	0	2.6	14.4	3.4	4.1	13.5	0	3.4	12.57	0	3.66	11.3	18.5	26.8
dw-m.en	45.3	1.32	2.82	24.6	0	6.8	18	0	7.9	10.1	0.5	18.3	16.86	0.39	15.29	7.2	0	14.4	58.8	0.5	3.2	12.6	6.2	5.8	16.7	0	5.7	21.99	0	6.28	13	21.2	31.6
dw-s.en	40.6	0.75	5.08	29.6	0.1	7.7	21.2	0.1	6.4	13.4	1.1	17.2	13.73	0	12.16	7.8	0	12.1	53.9	0.6	2.8	7.4	15.7	2.2	18	0.3	5.2	24.61	0	4.19	13	20.5	29.8
sm4t-l	41.73	0	4.32	17.7	0.3	2.4	20.3	0	5.1	7.9	1.7	13.9	7.06	1.96	9.02	6.4	0	14.8	29.6	1.6	4.2	1.2	21.8	0.2	20.2	0	2.3	7.33	0	2.09	4.6	15.3	25.3
sm4t-m	39.29	0	3.76	21.9	0.4	3.7	24.5	0.1	6.9	8.5	1.9	15.6	15.69	1.18	3.53	5.1	0	15.2	47.2	1	3.2	3.7	17.2	2.5	23.8	0	3.6	9.95	0	3.66	5.8	16.4	29
hubert	66.17	3.2	0.56	58.4	0.2	4.4	21.8	0.1	3.1	62.7	1.3	10.4	0	0	0	47.2	0	8.6	91.4	0.2	0.5	5.3	52.4	0.1	17.2	0	1	56.54	0	5.76	52.3	7.3	30.5
sm4t-v2-l	42.29	0	3.38	15	0.2	3	17.9	0.2	2.6	9.1	2	16.2	13.33	0.39	2.75	4.1	0	17.6	49.1	1.6	4.9	2.8	20.6	0.9	17.4	0	2.2	6.28	0	2.62	8.9	13.8	26.4
spllm-1.5B	50.56	1.13	1.32	39.9	0.9	6.7	42.9	0.3	4.5	24.4	6.2	20.7	3.14	1.57	1.96	6.67	0	17.5	55.51	3.81	1	1.3	2	0.3	38.8	0.5	3.5	29.32	0	2.62	22.3	13.82	33.6
w2v2-large	72.74	0.19	0.75	64.6	0.2	3.5	46.6	0.1	9.7	55.3	1.2	14.1	1.96	2.35	0	50.2	0	7.8	87.2	0	0.2	6.7	36	0	35.5	0	8.8	58.12	0	3.66	43.7	9.4	37.7
w-large	33.83	0.56	1.69	14.9	0	2.2	12.1	0	2.7	6.2	0.3	10.6	20	0.39	7.45	6.4	0	9.9	48.6	0	1.3	12.5	5.8	2.5	11.9	0	2.4	7.85	0	1.05	8.5	14.1	23.7
w-l-v2	36.84	0	2.63	14.3	0	2.4	11.3	0	2.1	5.1	1.1	10.2	16.86	0	9.8	6.2	0	11.2	45.5	0.2	1	10.3	3.4	1.7	12.1	0	2.7	6.28	0	2.09	6.7	15.1	23.4
w-l-v3	31.02	0	2.63	10.3	0	1.9	9	0	2.4	6.5	0.1	8.3	16.86	0	7.06	6.5	0	10.4	49.8	0	1.1	11.3	9	2.6	9.1	0	1.8	5.24	0	1.57	6.6	17.9	25.2
w-l-v3-t	30.83	0.38	2.82	13.5	0	2.6	10.4	0	2.2	5.1	0.6	8.9	21.57	0	9.41	6.4	0	11.3	54.9	0	1.6	13.2	12.8	2	10	0.1	2.2	10.99	0	3.14	7.5	17.1	23.9
w-m	34.59	0.19	1.88	17	0.1	3.4	14.8	0	2.9	6.8	0.7	10.4	16.86	0.39	10.98	6.5	0	11.2	51.5	0	1.9	11.2	6.6	1.9	13.8	0	2	10.47	0	2.09	10.1	17.2	24
w-m.en	32.71	0.19	2.63	16.4	0	2.8	12.8	0	3.1	6.6	0.4	10.1	17.25	0.39	9.02	6.6	0	10.3	50.1	0.8	2.4	10.2	9.4	3.5	12.5	0	2.6	15.71	0	2.09	6.6	16.1	25.7
w-m	38.16	0	2.63	28.5	0	4.8	19.6	0	5.7	10.9	0.7	14.3	12.55	0.78	7.45	6.6	0	12	54.3	0.3	1.9	7.7	8.3	2.2	17.5	0	3.8	17.28	0	1.05	10.2	18.3	27.6
w-s.en	37.22	0.38	3.38	27	0	5.4	18.2	0	3.9	8.2	1	14.8	12.94	0.78	6.27	6.9	0	11.8	57.9	0.1	1.4	6.9	11.3	2.1	15.2	0	3.4	16.23	0	3.14	9.2	19.4	26.2
w-tiny	36.28	2.07	4.89	26.5	0.3	10.3	33.7	0	16.8	15.3	1	26.6	6.27	0.78	6.67	9.5	0	16.9	32.4	1.8	2.5	0	18.2	0.5	32	0	12.6	33.51	0	9.95	15.6	23.7	40.3
w-tiny.en	34.59	2.26	4.32	29.2	0.3	12.8	30.6	0.1	14.2	12.8	1	21.4	9.02	0.78	5.1	9.3	0	14.5	33	1.7	3.2	0.4	25.9	1.1	25.7	0	9.6	34.03	0	6.28	13.5	22	36.5

Table 9: Non-Hallucination error analysis across various datasets and models.The table shows the percentage of Phonetic (P), Oscillation (O), and Language (L) errors for each model evaluated on different datasets. Abbreviations. w – whisper, s – small, m – medium, l – large, t – turbo, dw – distil-whisper, sm4t – seamless, w2v2 – wav2vec2, spllm – SpeechLLM, Qwen2 – Q2A - Qwen2-Audio, SC - Supreme Court.

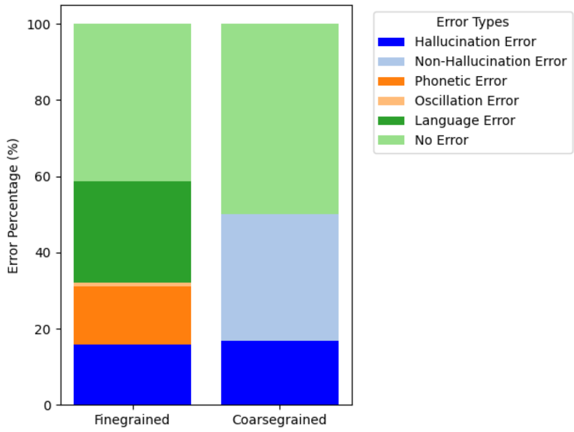


Figure 8: Finegrained vs coarsegrained error rate distribution averaged across all models and datasets.

Attribute	Value
Reference	lufthansa four three nine three descend to flight level two seven zero
Transcription	Lufthansa 4393, descent flight level 270.
WER	75.0
Hallucination	No Error

Table 10: WER and error category labeled by LLMs for whisper-medium.