Detecting Hallucinations in SpeechLLMs at Inference Time Using Attention Maps

Jonas Waldendorf[1,*] **Bashar Awwad Shiekh Hasan**[2] **Evgenii Tsymbalov**[2]
[1]University of Edinburgh [2]Amazon AGI
jonas.waldendorf@ed.ac.uk etsymba@amazon.de
*Work completed during an internship at Amazon AGI

Abstract

Hallucinations in Speech Large Language Models (SpeechLLMs) pose significant risks, yet existing detection methods typically rely on gold-standard outputs that are costly or impractical to obtain. Moreover, hallucination detection methods developed for text-based LLMs do not directly capture audio-specific signals. We investigate four attention-derived metrics: AUDIORATIO, AUDIOCONSISTENCY, AUDIOENTROPY, and TEXTENTROPY, designed to capture pathological attention patterns associated with hallucination, and train lightweight logistic regression classifiers on these features for efficient inference-time detection.

Across automatic speech recognition and speech-to-text translation tasks, evaluations on Qwen-2-Audio and Voxtral-3B show that our approach outperforms uncertainty-based and prior attention-based baselines on in-domain data, achieving improvements of up to +0.23 PR-AUC, and generalises to out-of-domain ASR settings. We further find that strong performance can be achieved with approximately 100 attention heads, improving out-of-domain generalisation compared to using all heads. While effectiveness is model-dependent and task-specific training is required, our results demonstrate that attention patterns provide a valuable tool for hallucination detection in SpeechLLMs.

1 Introduction

Speech is a vital modality for many modern technologies, including personal voice assistants (Hoy, 2018), automatic transcription services (Radford et al., 2023), and audio translation systems (Barrault et al., 2023). Despite recent advances, speech recognition and translation models can produce hallucinated content, fluent and plausible outputs that are not grounded in the input audio. While typical transcription errors often preserve semantic meaning and may still be interpretable by downstream users, hallucinations can introduce fabricated or misleading information, leading to critical failures. Consequently, identifying hallucinations and models that are predisposed to hallucinate remains an important area of research.

Hallucinations are pathological generations that are not grounded in the input and instead rely on distributional patterns learned from training data. Such outputs are often fluent, yet contain fabrications that alter the semantic content relative to the source audio (Koudounas et al., 2025; Atwany et al., 2025). Hallucination detection has been extensively studied in the text modality, for example, in retrieval-augmented generation (Shuster et al., 2021) and machine translation (Guerreiro et al., 2023a). In contrast, existing approaches for SpeechLLMs primarily compare model hypotheses against gold-standard outputs to identify hallucinations (Frieske and Shi, 2024; Koudounas et al., 2025). While effective, such supervised methods require costly data annotation and often depend on external models to perform the comparison.

Lightweight detection methods that leverage the internal representations of SpeechLLMs offer several advantages. First, they can be deployed at inference time for online filtering, preventing harmful outputs at the source, for example, by rerouting problematic inputs. Second, they enable low-cost flagging of pathological generations for offline analysis. Third, they can be combined with complementary signals such as uncertainty estimation (UE) metrics to capture a broader range of errors. Motivated by these benefits, we propose training lightweight classifiers on internal SpeechLLM representations to detect hallucinations at inference time.

Our approach exploits patterns in attention heads as signals for hallucination detection. The underlying intuition is that when models generate outputs that are not grounded in the input, their attention exhibits distinctive patterns that can be detected

Findings of the Association for Computational Linguistics: ACL 2026, pages 43276–43289

automatically. Recent work has demonstrated the effectiveness of attention-based signals for hallucination detection in text LLMs (Chuang et al., 2024; Vazhentsev et al., 2025; Sriramanan et al., 2024a). However, these methods have not been adapted to SpeechLLMs, where the input modality introduces fundamentally different attention dynamics. Audio representations are substantially longer than text, and the alignment between input frames and output tokens differs from text-to-text generation.

We address this gap by analysing attention patterns in two SpeechLLMs, `Qwen-2-Audio` (Chu et al., 2024) and `Voxtral-3B` (Liu et al., 2025), and training logistic regression classifiers on audio-specific attention features for hallucination detection.

Contributions

- We propose four audio-focused attention metrics, AUDIORATIO, AUDIOCONSISTENCY, AUDIOENTROPY, and TEXTENTROPY, designed to capture hallucination-related attention patterns in SpeechLLMs.
- We develop lightweight logistic regression classifiers trained on these attention-derived features. Our models outperform uncertainty estimation methods and existing attention-based baselines for hallucination detection, achieving improvements of up to +0.23 PR-AUC on in-domain ASR data.
- We show that strong detection performance can be achieved using approximately 100 attention heads. This improves out-of-domain generalisation compared to using all heads, and we further demonstrate that effectiveness varies across models and tasks.

2 Related Work

Uncertainty Estimation. A common approach to hallucination detection relies on UE metrics. Metrics derived from the LLM logits provide a low-cost and effective signal for identifying when a model is uncertain in its generation, which in turn correlates strongly with hallucinated content (Huang et al., 2024; Vashurin et al., 2025; Vazhentsev et al., 2025). Metrics such as SEQLOGPROB (the log probability of the entire output sequence), PERPLEXITY, and MEAN ENTROPY (defined as the average token-level entropy) have all been shown to be effective for hallucination detection (Malinin and Gales, 2021; Guerreiro et al., 2023b).

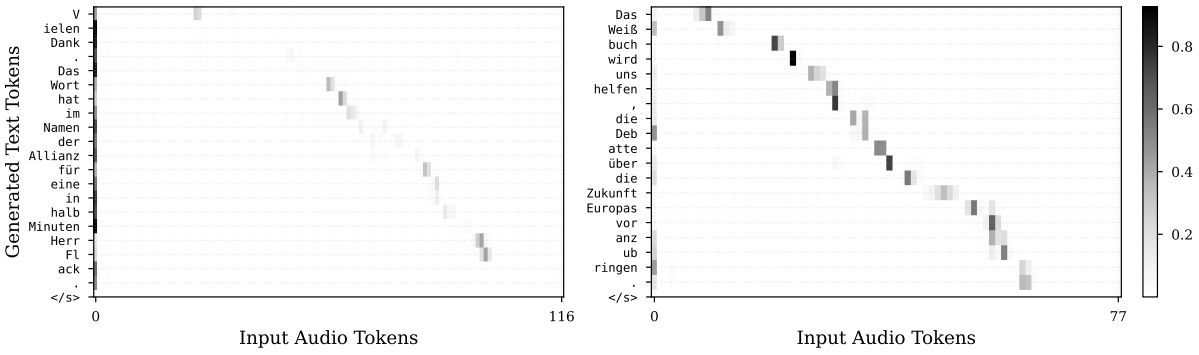
Hallucination Detection via Attention. Several approaches exploit attention patterns to detect hallucinations in text LLMs. Vazhentsev et al. (2025) analyse causal attention to the previous token, while Sriramanan et al. (2024b) average attention maps across layers. LOOKBACK-LENS (Chuang et al., 2024) train a classifier on the ratio of attention allocated to the input versus the auto-regressive prefix. We build on this line of work by adapting attention-based detection to SpeechLLMs using four audio-focused attention metrics.

Related work in machine translation has shown that hallucinations correlate with reduced diagonal entropy in attention maps (Voita et al., 2021; Raunak et al., 2021). We incorporate entropy-based features in our approach, but compute them specifically over audio attention, reflecting the substantial length disparity between audio frame sequences and output tokens.

Reference-Based Methods. An alternative class of approaches relies on gold-standard outputs to detect hallucinations. Frieske and Shi (2024) study hallucinations in ASR systems and identify them using a combination of semantic similarity between hypotheses and references, together with output fluency under a language model. Koudounas et al. (2025) introduce SHALLOW, a benchmark that categorises hallucinations into four types: lexical fabrications, phonetic fabrications, morphological hallucinations, and semantic hallucinations. While such reference-based methods enable fine-grained analysis, they require access to reference transcriptions that are often unavailable in deployment settings. This limitation motivates our reference-free detection approach.

3 Methodology

Attention to Input in SpeechLLMs. Based on recent findings (Vazhentsev et al., 2025; Wang et al., 2025), we hypothesise that certain attention heads exhibit distinctive patterns whilst generating hallucinated content, which can be exploited for detection at inference time. We focus on two such patterns. The first consists of diagonal attention structures that encode a temporal relationship between the audio input and the generated text. These patterns often degrade when the model is uncertain,



(a) (b)

Figure 1: Attention to audio tokens for Layer 25, Head 30 in `Voxtral-3B`. (a) Hallucination: attention collapses to early audio frames, losing temporal alignment. (b) Correct transcription: diagonal pattern reflects alignment between audio input and generated text.

with attention falling back to the early portion of the audio input. The second pattern involves heads that balance attention between the audio input and the auto-regressive text prefix.

Figure 1 shows two attention maps from Layer 25, Head 30 of `Voxtral-3B`, comparing a hallucinated transcription with a correct one. The figure illustrates two effects. First, the diagonal attention pattern, which captures the alignment between the audio input and the generated text, degrades during hallucination. Second, attention falls back to the beginning of the audio input rather than shifting toward the text prompt or the auto-regressive prefix.

3.1 Metrics

We experiment with four metrics designed to capture characteristic features of attention heads. Consider a SpeechLLM with L layers and H heads, an input sequence

$$X = \{x_{a,1}, x_{a,2}, \ldots, x_{a,N} \mid x_{t,1}, x_{t,2}, \ldots, x_{t,M}\}$$

where $x_{a,i}$ denotes an audio input token and $x_{t,j}$ denotes a text input token, and an output sequence

$$Y = \{y_1, y_2, \ldots, y_{t-1}\}$$

We define each metric at a given decoding step t and then describe a shared aggregation strategy.

AUDIORATIO: The ratio of attention allocated to audio input tokens versus auto-regressive text tokens.

For AUDIORATIO at decoding step t and head (l, h), where ART denotes the auto-regressive text prefix, we compute

$$A_t^{l,h}(\text{Audio}) = \sum_{i=1}^{N} a_{t,i}^{l,h}, \tag{1}$$

$$A_t^{l,h}(\text{ART}) = \sum_{j=N+M+1}^{N+M+t-1} a_{t,j}^{l,h}. \tag{2}$$

The audio ratio is then defined as

$$AR_t^{l,h} = \frac{A_t^{l,h}(\text{Audio})}{A_t^{l,h}(\text{Audio}) + A_t^{l,h}(\text{ART})}. \tag{3}$$

This metric builds on LOOKBACK-LENS by computing the ratio of attention allocated to input and output tokens, but restricts the input side to audio tokens only, since the text prompt in our tasks contains only instructions.

AUDIOCONSISTENCY: The Pearson correlation coefficient between audio attention vectors at consecutive decoding steps. This metric is defined only after the first generation step and is computed as

$$AC_t^{l,h} = r\left(a_{1:N}^{l,h,t}, a_{1:N}^{l,h,t-1}\right) \tag{4}$$

As illustrated in Figure 1, this metric aims to capture attention fallback behaviour. When hallucinating, the model often focuses attention at the beginning of the audio input, thereby increasing the similarity between consecutive attention distributions. An important limitation of AUDIOCONSISTENCY is that it explicitly targets heads with diagonal attention patterns. Some heads may attend strongly to early audio positions even during correct generations.

Both entropy-based metrics are computed by selecting the relevant attention weights at decoding step t, re-normalising them, and computing entropy as

$$AE_t^{l,h} = H\left(\frac{a_{1:N}^{l,h,t}}{\sum_{i=1}^{N} a_i^{l,h,t}}\right). \quad (5)$$

AUDIOENTROPY: The entropy of re-normalised attention weights over audio input tokens. AUDIOENTROPY aims to capture uncertainty in the audio input and provides a useful signal even for heads that do not exhibit clear diagonal attention patterns.

TEXTENTROPY: The entropy of re-normalised attention weights over text input tokens. The goal of this metric is to capture uncertainty in text-focused attention heads.

For each metric, values are computed at every decoding step and then averaged across all time steps to obtain a single value per layer and head. These values are concatenated into a feature vector of dimension $L \times H$ for each metric. We then use these vectors as input features to train logistic regression models as hallucination detectors.

4 Experimental Setup

We evaluate our models across two speech-related tasks: ASR (automatic speech recognition) and S2TT (speech-to-text translation).

4.1 Evaluation Datasets

For ASR, we evaluate on two datasets: VOXPOPULI[1] Wang et al. (2021) and CALLHOME[2] Canavan et al. (1997). For VOXPOPULI, we experiment with the English, German, French, and Spanish language splits. When evaluating hallucination detection, we combine all VOXPOPULI test sets for a total of 7,080 sentences. We pre-process CALLHOME by removing sentences shorter than two words (tokenising on whitespace), leaving 3,916 examples.

For S2TT, we evaluate on the English, German, French, and Spanish subsets of the FLEURS[3] Conneau et al. (2022) multilingual speech translation dataset, totalling 4,613 examples.

[1] https://huggingface.co/datasets/facebook/voxpopuli

[2] https://catalog.ldc.upenn.edu/LDC97S42

[3] https://huggingface.co/datasets/google/fleurs

4.2 Labelling Hallucinations

Training and evaluating hallucination detection models requires binary labels. Since manual annotation is expensive, we first collect a small set of human annotations to calibrate automatic labelling thresholds, and then apply these thresholds to label the remaining data automatically.

Human Annotation. We manually annotated 1,950 examples sampled from the English and German VOXPOPULI development sets, identifying 142 hallucinated outputs. An output was labelled as a hallucination if it contained fluent but fabricated content that was not grounded in the input audio.

Dataset	Language	WER	SHS	Hal.%
VOXPOPULI	De	13.09	11.47	5.7
	En	7.01	8.29	1.1
	Es	9.47	11.61	2.6
	Fr	10.98	10.73	3.6
CALLHOME	En	20.90	19.03	20.6

Table 1: WER, SHS, and hallucination percentage for `Qwen-2-Audio` on VOXPOPULI and CALLHOME test sets.

Automatic Labelling. To label hallucinations automatically, we threshold on a combination of lexical and semantic information. This approach follows prior findings that semantic content is critical for identifying ASR hallucinations Frieske and Shi (2024). We use WER (word error rate) to capture surface-level errors and the SEMANTIC HALLUCINATION SCORE (SHS) from the SHALLOW benchmark (Koudounas et al., 2025) to capture semantic divergence (Appendix C):

$$\text{Hallucination} = \mathbb{I}[\text{WER} + \text{SHS} > 0.7], \quad (6)$$

where $\mathbb{I}[\cdot]$ denotes the indicator function.

Threshold Selection. The threshold was tuned using stratified five-fold cross-validation on the human-annotated subset. We prioritised high precision (0.979) in order to obtain a clean training signal, accepting lower recall (0.443) as a trade-off.

Tables 1 and 2 report hallucination rates for both models across evaluation datasets. Hallucination rates are consistently low on VOXPOPULI test sets, but increase substantially, up to 20%, on the noisier

Dataset	Language	WER	SHS	Hal.%
VOXPOPULI	De	11.66	11.59	6.3
	En	7.16	8.76	1.1
	Es	8.59	11.27	2.5
	Fr	10.46	10.60	3.3
CALLHOME	En	18.72	16.52	15.8

Table 2: WER, SHS, and hallucination percentage for `Voxtral-3B` on VOXPOPULI and CALLHOME test sets.

CALLHOME dataset. Both models exhibit similar trends, although `Voxtral-3B` generally achieves lower error rates, with the exception of German.

4.3 Training Data

We train all logistic regression models on 40,000 examples from the VOXPOPULI training data: 10,000 each for English, German, Spanish, and French. This language mix exposes the model to varying hallucination rates across languages (Tables 1 and 2), improving robustness to distribution shifts. Under this setup, `Qwen-2-Audio` produces 1,537 hallucinations (3.8%), while `Voxtral-3B` produces 1,178 hallucinations (2.9%).

For the S2TT task, we train on the FLEURS training set (16,776 examples), using hallucination labels derived from COMET scores, and evaluate on the held-out test set.

4.4 Logistic Regression Model Training

Using the attention-derived features described above, we train logistic regression models for hallucination detection. Model hyperparameters are reported in the Appendix (Table 9). We apply Min-Max scaling to AUDIOENTROPY and TEXTENTROPY to ensure that all feature values lie within the range $[0, 1]$.

We employ two feature selection strategies. First, we train an L2-regularised model and rank attention heads by the magnitude of their coefficients, scaled by the original feature standard deviation. Second, we train an L1-regularised model to perform feature pruning. Specifically, we run five-fold cross-validation and retain heads with non-zero coefficients in at least four of the five folds. We then retrain an L2-regularised model using only these retained heads. We refer to this variant as the *Stable Features* model.

4.5 Evaluation Metrics

ASR Evaluation. For ASR, we report F1-score, precision, recall, and PR-AUC, computed using the predicted probabilities from the logistic regression models.

In addition, we report the Prediction Rejection Ratio (PRR) (Malinin et al., 2017; Malinin and Gales, 2021), which measures the effectiveness of predicted probabilities in rejecting low-quality samples. Intuitively, PRR quantifies how closely probability-based rejection approaches oracle performance, where PRR $= 1$ corresponds to a perfect ordering with respect to a quality metric. We report PRR at 10% rejection for VOXPOPULI and at 30% rejection for CALLHOME, using SHS as the quality metric. These rejection rates reflect the approximate prevalence of hallucinations in each dataset. Further details are provided in Appendix D.

S2TT Evaluation. For S2TT, we use XCOMET-XL[4] (Guerreiro et al., 2024), which has been shown to correlate strongly with hallucinations in machine translation. Based on the empirical distribution of COMET scores, we label the bottom 5% of examples as hallucinations. We report F1-score, precision, recall, and PR-AUC, following the same protocol as for ASR. We also report PRR@10%, using COMET as the quality metric.

4.6 Baselines

As baselines, we consider simple UE metrics, namely MEAN ENTROPY (the mean token-level entropy) and PERPLEXITY, which have been shown to correlate with hallucinations in text summarisation and machine translation. In addition, we include two attention-based baselines, RAUQ (ENTROPY) and ATTENTIONSCORE, as attention-based methods for hallucination detection.

5 Results

5.1 ASR Results

Tables 3 and 4 report hallucination detection performance for `Qwen-2-Audio` and `Voxtral-3B`, respectively, on the VOXPOPULI and CALLHOME test sets. We present results for logistic regression models trained using the following feature configurations.

Combined concatenates all four attention metrics across all layers and heads, yielding $L \times H$ features per metric. **AUDIORATIO Only** uses only the

[4] https://huggingface.co/Unbabel/XCOMET-XL

AUDIORATIO metric across all layers and heads, which was the best-performing single metric on the VOXPOPULI validation set. **Top N** selects the top N layer–head pairs per metric based on coefficient magnitude, with N tuned separately for each model on the validation set.

Dataset	Method	Hal.%	Acc	F1	Prec	Rec	PR-AUC	PRR@k
BASELINES								
VOXPOPULI								
	Mean Entropy	3.33	0.97	0.50	0.50	0.50	0.49	0.43
	Perplexity	3.18	0.97	0.50	0.51	0.49	0.49	0.43
	RAUQ Entropy	3.42	0.96	0.46	0.45	0.47	0.47	0.46
	Attention Score	1.23	0.96	0.08	0.15	0.06	0.04	0.30
	Random	50.27	0.50	0.06	0.03	0.50	–	–
CALLHOME								
	Mean Entropy	38.23	0.76	0.58	0.45	0.83	0.67	0.59
	Perplexity	37.89	0.76	0.58	0.45	0.83	0.69	0.61
	RAUQ Entropy	24.40	0.80	0.56	0.52	0.61	0.61	0.54
	Attention Score	0.00	0.79	0.00	0.00	0.00	0.14	-0.13
	Random	50.03	0.50	0.29	0.20	0.49	–	–
LOGISTIC REGRESSION								
Combined (4096 features)								
VOXPOPULI	LR	3.62	0.97	0.55	0.52	0.57	0.58	0.51
CALLHOME	LR	75.99	0.43	0.41	0.26	0.96	0.61	0.53
AUDIORATIO *(1024 features)*								
VOXPOPULI	LR	3.57	0.97	0.56	0.54	0.58	0.56	0.49
CALLHOME	LR	66.97	0.52	0.45	0.29	0.95	0.60	0.53
Top 75 (300 features)								
VOXPOPULI	LR	3.87	0.97	0.52	0.49	0.57	0.58	0.49
CALLHOME	LR	79.05	0.40	0.40	0.25	0.97	0.59	0.50

Table 3: Hallucination detection performance (F1, Precision, Recall, PR-AUC, and PRR@k) for `Qwen-2-Audio` on VOXPOPULI and CALLHOME test sets, with k=10% and k=30% respectively.

Attention-based features outperform baselines on in-domain data. On the VOXPOPULI test set, logistic regression models consistently outperform uncertainty-based baselines for both SpeechLLMs. For `Qwen-2-Audio`, the strongest baseline, MEAN ENTROPY, achieves an F1-score of 0.50 and a PR-AUC of 0.49, whereas logistic regression attains up to 0.56 F1 and 0.58 PR-AUC. The improvement is substantially larger for `Voxtral-3B`. Here, MEAN ENTROPY reaches 0.42 F1 and 0.44 PR-AUC, while AUDIORATIO achieves 0.64 F1 and 0.67 PR-AUC, corresponding to a gain of 0.23 PR-AUC.

Label-free evaluation supports these results. PRR@k improves by 0.05 for `Qwen-2-Audio` and by 0.13 for `Voxtral-3B` relative to the strongest baseline that does not use attention. Together, these findings demonstrate that attention-based features provide a strong discriminative signal for hallucination detection on in-domain data. Qualitative examples are provided in Appendix E.

Uncertainty baselines excel on noisy speech. Baseline performance is notably stronger on CALLHOME than on VOXPOPULI for both models. For `Qwen-2-Audio`, PERPLEXITY achieves a PR-AUC of 0.69 on CALLHOME compared to 0.49 on VOXPOPULI, while for `Voxtral-3B`, MEAN ENTROPY achieves 0.59 versus 0.44. We attribute this effect to the inherently noisy nature of CALLHOME. Conversational speech with overlapping speakers and frequent disfluencies induces higher model uncertainty, making uncertainty-based metrics more effective. This interpretation is supported by the average MEAN ENTROPY values, which are approximately 0.10 for VOXPOPULI and 0.30 for CALLHOME. We observe a consistent pattern across models: `Voxtral-3B` achieves stronger overall ASR performance, while uncertainty-based methods are comparatively more effective for `Qwen-2-Audio`. In contrast, our attention-based approach performs best for stronger models, cleaner data, and settings in which hallucinations are relatively rare.

Dataset	Method	Hal.%	Acc	F1	Prec	Rec	PR-AUC	PRR@k
BASELINES								
VOXPOPULI								
	Mean Entropy	2.10	0.97	0.42	0.55	0.34	0.44	0.43
	Perplexity	3.12	0.96	0.40	0.42	0.39	0.41	0.40
	RAUQ Entropy	2.03	0.96	0.35	0.47	0.28	0.32	0.43
	Attention Score	7.12	0.91	0.17	0.12	0.26	0.09	0.10
	Random	50.27	0.50	0.06	0.03	0.49	–	–
CALLHOME								
	Mean Entropy	15.83	0.86	0.55	0.55	0.55	0.59	0.57
	Perplexity	25.18	0.79	0.50	0.40	0.64	0.56	0.54
	RAUQ Entropy	11.56	0.86	0.51	0.60	0.44	0.52	0.53
	Attention Score	53.92	0.46	0.23	0.15	0.51	0.15	-0.01
	Random	50.03	0.49	0.22	0.15	0.46	–	–
LOGISTIC REGRESSION								
Combined (3840 features)								
VOXPOPULI	LR	3.23	0.98	0.64	0.65	0.62	0.69	0.56
CALLHOME	LR	32.80	0.76	0.50	0.37	0.77	0.55	0.52
AUDIORATIO *(960 features)*								
VOXPOPULI	LR	3.04	0.98	0.64	0.68	0.61	0.67	0.54
CALLHOME	LR	29.13	0.79	0.52	0.40	0.74	0.58	0.55
Top 75 (300 features)								
VOXPOPULI	LR	2.99	0.98	0.62	0.66	0.58	0.68	0.55
CALLHOME	LR	27.26	0.80	0.55	0.43	0.74	0.61	0.57

Table 4: Hallucination detection performance (F1, Precision, Recall, PR-AUC, and PRR@k) for `Voxtral-3B` on VOXPOPULI and CALLHOME test sets, with k=10% and k=30% respectively.

Out-of-domain generalization is model-dependent. Logistic regression performance on CALLHOME differs markedly between the two models. For `Qwen-2-Audio`, logistic regression predicts a large number of false positives, with predicted hallucination rates ranging from 66.97% to 79.05%, compared to an actual rate of 20.90%. As a result, the best logistic regression model achieves an F1-score of only 0.45 using AUDIORATIO, compared to 0.58 for both PERPLEXITY and MEAN ENTROPY. PRR@k similarly drops by 0.11 relative to the strongest baseline for the Top

75 configuration.

In contrast, `Voxtral-3B` maintains competitive out-of-domain performance. The Top 75 model achieves an F1-score of 0.55 and a PR-AUC of 0.61 on CALLHOME, matching or exceeding MEAN ENTROPY. Notably, PRR@k is higher on CALLHOME than on VOXPOPULI for `Voxtral-3B` in all configurations except the Combined model. This suggests that, for `Voxtral-3B`, attention-based features generalise well despite the absence of comparable training data. Overall, these results indicate that out-of-domain performance benefits from combining multiple attention head metrics, but remains strongly model-dependent.

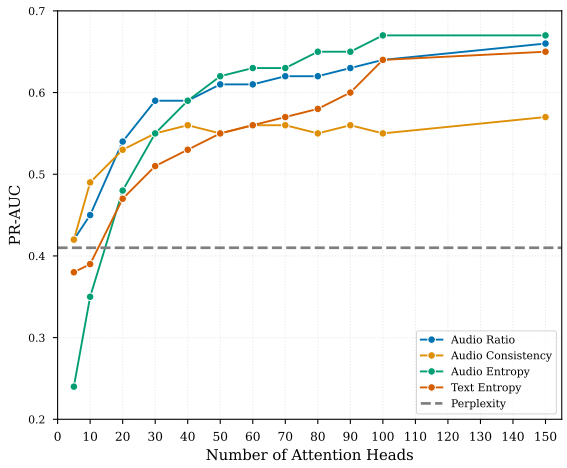


Figure 2: PR-AUC as a function of feature count for `Voxtral-3B` on VOXPOPULI. Each curve represents a different attention metric, with heads ranked independently by coefficient magnitude using L2 logistic regression.

`Qwen-2-Audio` performance gaps reflect threshold sensitivity. Despite `Qwen-2-Audio` 's low F1-scores on CALLHOME, PR-AUC remains comparable to VOXPOPULI, ranging from 0.61 to 0.62 on CALLHOME versus 0.56 to 0.58 on VOXPOPULI. This indicates that the underlying attention features still contain discriminative information. The elevated false positive rate appears to be driven primarily by threshold miscalibration under distribution shift, rather than by fundamentally different attention behaviour. PRR@k results also support that the gap between datasets is substantially smaller than that suggested by the F1-scores. The label-free metric confirms that the proposed approach can still prioritise semantically problematic outputs even when binary decision thresholds are poorly calibrated.

Fewer features improve out-of-domain generalization. Figure 2 plots PR-AUC as a function of feature count for `Voxtral-3B` for VOXPOPULI. Performance improvements plateau after approximately 100 attention heads per metric. With as few as five features, all metrics perform similarly to, or worse than, the PERPLEXITY baseline, indicating that while individual heads may carry signal, combining multiple heads is necessary for robust performance. AUDIOCONSISTENCY performs well with relatively few heads but saturates earlier, which is expected given that it targets specific diagonal attention patterns. In contrast, AUDIOENTROPY requires aggregating signals from approximately 30 or more heads to reach comparable PR-AUC, after which performance scales more smoothly.

Table 5 further shows that reducing the number of features trades a small amount of in-domain performance for improved generalisation. The *Stable Features* model (see Appendix: Table 10), which uses 99 features, shows a reduction of 0.02 PR-AUC on VOXPOPULI compared to Top 75, which uses 300 features, but achieves the highest PR-AUC of any logistic regression model on CALLHOME at 0.64. Using only AUDIORATIO features with an equivalent feature budget identifies fewer hallucinations, suggesting that combining metrics becomes increasingly important as the feature count is reduced. Analysis of the selected features shows that the L1-regularised model prioritises AUDIORATIO and AUDIOCONSISTENCY heads, as reported in Appendix 10. This observation is consistent with the scaling trends shown in Figure 2.

Dataset	Method	Hal.%	Acc	F1	Prec	Rec	PR-AUC	PRR@k
Top 25 (100 features)								
VOXPOPULI	LR	2.98	0.97	0.60	0.64	0.56	0.65	0.53
CALLHOME	LR	25.08	0.81	0.54	0.44	0.70	0.61	0.56
AUDIORATIO *Only (100 features)*								
VOXPOPULI	LR	2.78	0.98	0.60	0.67	0.55	0.64	0.49
CALLHOME	LR	21.44	0.83	0.55	0.47	0.64	0.58	0.58
Stable Features (99 features, threshold=0.8)								
VOXPOPULI	LR	2.80	0.98	0.60	0.66	0.55	0.66	0.52
CALLHOME	LR	29.96	0.79	0.53	0.41	0.77	0.64	0.58

Table 5: Hallucination detection performance (F1, Precision, Recall, PR-AUC, and PRR@k) for `Voxtral-3B` with reduced feature sets on VOXPOPULI and CALLHOME test sets, with k=10% and k=30% respectively.

5.2 S2TT Results

Training on ASR data does not generalise to S2TT. We first evaluate whether logistic regression models trained on ASR data can transfer to the S2TT task. Using the best-performing ASR

configurations, namely Top 75 logistic regression for both `Qwen-2-Audio` and `Voxtral-3B`, we observe PR-AUC scores of only 0.15 and 0.08, respectively, which is only marginally above random performance. PRR@k metrics corroborate this result, with values of 0.30 for `Qwen-2-Audio` and 0.16 for `Voxtral-3B`, compared to baseline PRR@k scores of 0.46 and 0.44, confirming that attention-based hallucination detectors trained on ASR data do not generalise to S2TT. This raises the question of whether attention patterns differ fundamentally between tasks, or whether the use of different labelling schemes, WER and SHS for ASR versus COMET for S2TT, necessitates task-specific classifiers.

Method	Hal.%	Acc	F1	Prec	Rec	PR-AUC	PRR@k
TRAINED ON S2TT DATA							
Baselines							
Mean Entropy	0.21	0.95	0.08	1.00	0.04	0.25	0.45
Perplexity	0.23	0.95	0.08	0.90	0.04	0.23	0.39
RAUQ Entropy	0.18	0.95	0.07	1.00	0.04	0.25	0.46
Attention Score	14.06	0.84	0.14	0.10	0.27	0.08	0.15
Random	50.99	0.49	0.09	0.05	0.50	–	–
Combined (4096 features)							
LR	1.32	0.96	0.29	0.69	0.18	0.43	0.67
AUDIORATIO *(1024 features)*							
LR	1.30	0.95	0.23	0.56	0.15	0.39	0.64
Top 150 (600 features)							
LR	1.51	0.95	0.28	0.61	0.18	0.44	0.67
TRAINED ON ASR DATA							
Top 75 (300 features)							
LR	89.66	0.14	0.09	0.05	0.86	0.15	0.30

Table 6: Hallucination detection performance (F1, Precision, Recall, PR-AUC, and PRR@k) for `Qwen-2-Audio` on the FLEURS S2TT test set, with k=10%.

Baselines struggle on S2TT. Baseline performance on S2TT is substantially weaker than on ASR. The strongest baseline achieves a PR-AUC of only 0.25 for `Qwen-2-Audio` and 0.17 for `Voxtral-3B`, compared to 0.49 and 0.44, respectively, on the VOXPOPULI ASR test set. This suggests that uncertainty-based metrics, while effective for ASR, are less suited to detecting hallucinations in the speech-to-text translation setting. This finding is notable, given that such metrics are often strong baselines in text-based machine translation.

In-domain training yields substantial improvements over baselines. Training logistic regression models directly on S2TT data yields substantial performance gains over all baselines. For `Qwen-2-Audio`, the Top 150 configuration achieves a PR-AUC of 0.44 and an F1-score of 0.28, corresponding to improvements of 0.19 PR-AUC and 0.20 F1 over RAUQ Entropy. For `Voxtral-3B`, Top 300 reaches a PR-AUC of 0.44 and an F1-score of 0.37, improving on MEAN ENTROPY by 0.27 and 0.26, respectively. Label-free evaluation further supports these results. PRR@k scores reach 0.67 for `Qwen-2-Audio` and 0.68 for `Voxtral-3B`, indicating that logistic regression effectively prioritises low-quality translations. However, for both models, recall remains low, suggesting that while attention-based features reliably identify severe hallucinations, they are less sensitive to more subtle S2TT errors.

Method	Hal.%	Acc	F1	Prec	Rec	PR-AUC	PRR@k
TRAINED ON S2TT DATA							
Baselines							
Mean Entropy	1.13	0.95	0.11	0.31	0.07	0.17	0.32
Perplexity	0.17	0.95	0.05	0.75	0.03	0.15	0.27
RAUQ Entropy	1.15	0.95	0.13	0.34	0.08	0.16	0.35
Attention Score	83.57	0.20	0.10	0.05	0.90	0.06	0.02
Random	49.04	0.51	0.09	0.05	0.51	–	–
Combined (3840 features)							
LR	2.25	0.95	0.37	0.60	0.27	0.43	0.66
AUDIORATIO *(960 features)*							
LR	1.80	0.97	0.35	0.45	0.29	0.38	0.66
Top 150 (600 features)							
LR	2.51	0.95	0.37	0.55	0.28	0.44	0.68
TRAINED ON ASR DATA							
Top 75 (300 features)							
LR	83.92	0.20	0.10	0.05	0.92	0.08	0.16

Table 7: Hallucination detection performance (F1, Precision, Recall, PR-AUC, and PRR@k) for `Voxtral-3B` on the FLEURS S2TT test set, with k=10%.

Feature combination is important for S2TT. S2TT performance benefits from combining multiple attention-based metrics. For `Voxtral-3B`, using only AUDIORATIO features reduces the F1-score from 0.37 to 0.35 and the PR-AUC from 0.44 to 0.38, compared to the Top 300 configuration. `Qwen-2-Audio` exhibits a similar trend, with drops of 0.05 in both F1-score and PR-AUC when comparing AUDIORATIO to Top 150. These results support that the S2TT task relies on complementary signals captured by different attention metrics. Combining features is therefore particularly important for translation.

5.3 Task-specific attention heads dominate.

The observed lack of cross-task generalisation raises a fundamental question: does the logistic regression model rely on a universal set of attention heads, or is head selection inherently task-dependent? To investigate this, we compute the

intersection of the Top 50 most informative attention heads for each metric across ASR and S2TT, as reported in Table 8. We select the Top 50 heads because they capture most of the discriminative signal, as shown by the scaling behaviour in Figure 2.

Across all metrics and both models, we observe limited overlap between selected heads. Combined with poor cross-task transfer, this suggests that hallucination-related attention features are largely task-specific. However, the low overlap may also reflect feature collinearity, in which redundant heads lead the model to select different yet functionally similar features across tasks.

	Common Heads	
Metric	`Qwen-2-Audio`	`Voxtral-3B`
AUDIORATIO	22%	18%
AUDIOCONSISTENCY	32%	26%
AUDIOENTROPY	10%	8%
TEXTENTROPY	14%	14%

Table 8: Intersection of the top-50 most important attention heads for each metric for the ASR and S2TT tasks.

Among the four metrics, AUDIOCONSISTENCY shows the highest cross-task stability, followed by AUDIORATIO, consistent with earlier results showing strong performance even with few heads. In contrast, both entropy-based metrics exhibit substantially lower consistency. This may be because entropy-based regressors assign high weights to heads with weak alignment to their nominal input modality. For example, AUDIOENTROPY may select heads that primarily attend to the autoregressive text prefix rather than audio, making the resulting features more sensitive to task-specific noise.

6 Conclusions

We presented an attention-based approach to hallucination detection in SpeechLLMs. We introduced four audio-focused attention metrics, AUDIORATIO, AUDIOCONSISTENCY, AUDIOENTROPY, and TEXTENTROPY, and used them to train lightweight logistic regression classifiers. We evaluated the approach on two SpeechLLMs across automatic speech recognition and speech-to-text translation tasks.

Our method outperforms all baselines on in-domain ASR, achieving improvements of up to +0.23 PR-AUC on VOXPOPULI with `Voxtral-3B`, with particularly strong gains on cleaner data. Reducing the feature set from over 1,000 attention heads to approximately 100 yields comparable in-domain performance while improving out-of-domain generalisation, indicating that larger feature sets are prone to overfitting. Effectiveness varies across models: `Voxtral-3B` generalises well across settings, whereas `Qwen-2-Audio` exhibits weaker transfer to the noisier CALLHOME dataset.

Models trained on ASR data do not generalise to S2TT, highlighting that hallucination patterns and their associated attention signals are task-specific. However, training on S2TT data yields comparable improvements over baselines, demonstrating that the approach is effective when task-appropriate supervision is available. Future work could combine attention-based features with uncertainty estimation metrics to capture complementary signals, explore cross-model transfer, and extend the method to additional speech tasks such as speech summarisation or spoken dialogue systems.

7 Limitations.

Our automatic labelling strategy achieves high precision (0.979) but low recall (0.443), which means that many true hallucinations are excluded from training. Beyond threshold calibration, we do not perform direct human evaluation of the automatically labelled data. In addition, we focus exclusively on binary detection of severe hallucinations, rather than modelling finer-grained distinctions in hallucination type or severity.

Our approach does not generalise across tasks. Classifiers trained on ASR data do not generalise to S2TT, necessitating task-specific training data. Effectiveness is also model-dependent, with `Voxtral-3B` exhibiting more robust cross-domain generalisation than `Qwen-2-Audio`. Moreover, our evaluation is limited to two SpeechLLMs and four languages, which constrains the scope of our conclusions.

Finally, although the proposed method is lightweight compared to alternatives such as resampling or ensemble-based approaches, extracting attention patterns introduces additional inference-time computation and memory overhead. While this overhead remains modest relative to full SpeechLLM inference, it may still be relevant in latency-sensitive deployment settings.

References

Hanin Atwany, Abdul Waheed, Rita Singh, Monojit Choudhury, and Bhiksha Raj. 2025. Lost in transcription, found in distribution shift: Demystifying hallucination in speech foundation models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23181–23203, Vienna, Austria. Association for Computational Linguistics.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, and 1 others. 2023. Seamlessm4t: massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Alexandra Canavan, David Graff, and George Zipperlen. 1997. CALLHOME American English Speech. LDC97S42, Web Download.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1419–1436, Miami, Florida, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *arXiv preprint arXiv:2205.12446*.

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

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023a. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023b. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.

Matthew B Hoy. 2018. Alexa, siri, cortana, and more: an introduction to voice assistants. *Medical reference services quarterly*, 37(1):81–88.

Hsiu-Yuan Huang, Yutong Yang, Zhaoxi Zhang, Sanwoo Lee, and Yunfang Wu. 2024. A survey of uncertainty estimation in llms: Theory meets practice. *Preprint*, arXiv:2410.15326.

Alkis Koudounas, Moreno La Quatra, Manuel Giollo, Sabato Marco Siniscalchi, and Elena Baralis. 2025. Hallucination benchmark for speech foundation models. *Preprint*, arXiv:2510.16567.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2022. Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. *Preprint*. Publisher: Open Science Framework.

Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, Sanchit Gandhi, Soham Ghosh, Srijan Mishra, Thomas Foubert, Abhinav Rastogi, Adam Yang, Albert Q. Jiang, Alexandre Sablayrolles, Amélie Héliou, and 87 others. 2025. Voxtral. *Preprint*, arXiv:2507.13264.

Andrey Malinin and Mark John Francis Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Andrey Malinin, Anton Ragni, Kate Knill, and Mark Gales. 2017. Incorporating uncertainty into deep learning for spoken language assessment. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 45–50, Vancouver, Canada. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024a. Llm-check: Investigating detection of hallucinations in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 34188–34216. Curran Associates, Inc.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024b. LLM-check: Investigating detection of hallucinations in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025. Benchmarking uncertainty quantification methods for large language models with LM-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248.

Artem Vazhentsev, Lyudmila Rvanova, Gleb Kuzmin, Ekaterina Fadeeva, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, Mrinmaya Sachan, Preslav Nakov, and Artem Shelmanov. 2025. Uncertainty-aware attention heads: Efficient unsupervised uncertainty quantification for llms. *Preprint*, arXiv:2505.20045.

Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Analyzing the source and target contributions to predictions in neural machine translation. 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)*, pages 1126–1140, Online. Association for Computational Linguistics.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. 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)*, pages 993–1003, Online. Association for Computational Linguistics.

Yingzhi Wang, Anas Alhmoud, Saad Alsahly, Muhammad Alqurishi, and Mirco Ravanelli. 2025. Calm-whisper: Reduce whisper hallucination on non-speech by calming crazy heads down. *ArXiv*, abs/2505.12969.

A Logistic Regression

All logistic regression models are trained using `scikit-learn`[5] (Pedregosa et al., 2011).

Hyperparameter	**L2**	**L1**
Penalty	`L2`	`L1`
Max iterations	5000	5000
Class weights	1:2 (positive)	1:5 (positive)
C	1	0.005
Solver	`lbfgs`	`liblinear`

Table 9: Hyperparameters for L2 logistic regression and L1 logistic regression, the latter used for stable feature selection. These parameters were selected on the VOXPOPULI validation set.

Table 9 lists the hyperparameters used to train our hallucination detection models, selected on the VOXPOPULI validation set. Unless otherwise stated, these parameters are used for all logistic regression models.

Metric	Features
AUDIORATIO	39
AUDIOCONSISTENCY	36
AUDIOENTROPY	20
TEXTENTROPY	4
Total	99

Table 10: Distribution of stable features (threshold $\geq$ 0.8, selected in 4 out of 5 folds).

Table 10 summarises the stable features selected using L1 regularisation. These are the heads present in 4 out of 5 folds after training on the VOXPOPULI training data. The distribution shows that features are selected across all heads, with TEXTENTROPY being a clear outlier, contributing only 4 features.

B S2TT Results

Tables 11 and 12 report chrF (Popović, 2015) and COMET scores, computed using XCOMET-XL.

The results mirror those in Tables 3 and 4, with `Voxtral-3B` outperforming `Qwen-2-Audio` across both metrics and all datasets.

C Semantic Hallucination Score

A defining property of hallucinations is substantial semantic divergence between the gold standard transcription and the hypothesis. To capture this semantic change, we adopt the semantic hallucination score (SHS) from the SHALLOW ASR hallucination benchmark (Koudounas et al., 2025). SHS combines local and global semantic error metrics to capture both fine-grained and utterance-level inconsistencies between hypothesis and reference transcriptions.

Local semantic errors are computed using a multi-scale sliding window approach. Each hypothesis window is matched to reference windows via maximum cosine similarity of contextual embeddings, with higher weights assigned to smaller windows. This emphasises token-level distortions while remaining sensitive to phrase-level mismatches. Global semantic errors comprise two components. The first is semantic distance, computed as the inverse cosine similarity of sentence embeddings. The second is semantic coherence, which combines BERTScore with an entailment probability derived from a natural language inference model.

As our data spans multiple languages, we replace the original monolingual models with multilingual alternatives. We use `xlm-roberta-base` (Conneau et al., 2020) for local embeddings, `paraphrase-multilingual-MiniLM` (Reimers and Gurevych, 2019) for sentence embeddings, and

Language Direction	chrF	COMET
En→De	54.19	87.06
En→Es	47.42	82.26
En→Fr	49.72	76.74
De→En	60.29	90.33
Es→En	54.43	87.20
Fr→En	60.17	86.37

Table 11: S2TT performance for `Qwen-2-Audio` on FLEURS.

Language Direction	chrF	COMET
En→De	57.73	90.65
En→Es	51.09	88.69
En→Fr	59.23	84.13
De→En	64.90	94.42
Es→En	59.11	92.07
Fr→En	63.25	90.82

Table 12: S2TT performance for `Voxtral-3B` on FLEURS.

[5] https://scikit-learn.org/stable/index.html

Figure 3: Examples of classifications using Top 75 with `Voxtral-3B` outputs on the VOXPOPULI test set.

`mDeBERTa-v3-base-xnli` (Laurer et al., 2022) for natural language inference.

D PRR Metric

Equation 7 defines the predicted rejection ratio (PRR). For a corpus $D = \{(x_j, y_j)\}$, let $p_j = P(x_j, y_j^*)$ denote the logistic regression model's predicted probability that hypothesis y_j^* is a hallucination. The rejection curve plots the average quality $Q(y_j, y_j^*)$ of the remaining samples after rejecting those with $p_j < \alpha$.

$$\text{PRR} = \frac{\text{AUC}_{\text{prob}} - \text{AUC}_{\text{random}}}{\text{AUC}_{\text{oracle}} - \text{AUC}_{\text{random}}} \quad (7)$$

PRR measures how much average quality improves when rejecting samples based on predicted probability, relative to oracle rejection. A PRR of 1 indicates rejection in exact oracle order. We report PRR over the first 10% of the rejection curve for VOXPOPULI and the first 30% for CALLHOME, reflecting the proportion of hallucinations in each dataset.

E VOXPOPULI Examples

Figure 3 presents four examples of `Voxtral-3B` outputs classified using Top 75 on the VOXPOPULI test set. The first two are true positives that MEAN ENTROPY misclassifies as non-hallucinations. The first contains both substitutions and insertions, while the second exhibits hallucination through

substitution alone. We also include one false positive and one false negative to illustrate typical failure modes. Both cases lie close to the decision boundary of our automatic labelling pipeline, highlighting the inherent difficulty of the task. Notably, MEAN ENTROPY also misclassifies both examples.

F GPU Usage and AI Statement

Inference and labelling are performed on eight `A100-40GB` GPUs, processing approximately 4.5 samples per second. A single experimental run covers approximately 57,000 ASR sentences and 21,000 S2TT sentences, requiring around 38.5 GPU hours. Over development and evaluation, we conducted approximately six complete iterations, totalling around 230 GPU hours. Including XCOMET scoring and SHS computation, we estimate an upper bound of approximately 300 GPU hours.

Code for this project was partially written with the assistance of an internal coding assistant. Internal AI tools were also used to assist with the language and presentation of the paper.