Selective WER: Principled Word Error Rate Evaluation Under Selective Prediction for Long-Form ASR

Anonymous Author(s)

ABSTRACT

Standard Word Error Rate (WER) lacks a clear definition when a subset of hypothesized words is intentionally ignored based on word-level uncertainty in long-form automatic speech recognition (ASR). We propose a principled evaluation framework consisting of three complementary metrics: Selective WER (sWER), which treats abstentions as deletions and cannot be gamed; Abstention-Aware WER (aWER), which measures error rate over committed words only; and the Area Under the Risk-Coverage Curve (AURCC), which summarizes selective prediction quality across all operating points. Our framework extends Levenshtein alignment with a three-symbol hypothesis vocabulary and includes an oracle-informed decomposition to separate beneficial abstention from harmful abstention. Experiments on synthetic ASR data show that well-calibrated uncertainty scores achieve AURCC of 0.460 ± 0.041 compared to 0.583 ± 0.070 for random scores, and that uncertainty-based abstention at 0.789 coverage reduces aWER to 0.004 versus 0.160 for random abstention. We recommend a four-number reporting protocol (standard WER, sWER, aWER with coverage, and AURCC) for any ASR system employing selective prediction.

ACM Reference Format:
Anonymous Author(s). 2026. Selective WER: Principled Word Error Rate Evaluation Under Selective Prediction for Long-Form ASR. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

Word Error Rate (WER) is the de facto standard metric for evaluating automatic speech recognition (ASR) systems [7]. It computes the minimum edit distance between a reference transcript and a hypothesis, counting substitutions, deletions, and insertions, normalized by the number of reference words [6]. Despite its simplicity and widespread adoption, WER assumes that the ASR system produces a *complete* hypothesis for every input utterance.

Recent advances in uncertainty estimation for neural ASR models [8, 9] have enabled *selective prediction*: the system can flag individual words it is uncertain about and abstain from committing to those predictions. This is particularly valuable in long-form ASR applications such as lecture transcription and interview processing, where errors in critical content words can significantly impact downstream usability.

However, as noted by Bondarenko et al. [1], it is not clear how to evaluate WER when some words are ignored in long-form speech recognition. The authors explicitly adopt alternative metrics—uncertainty ratio and recall of error detection—precisely because WER under selective prediction is ill-defined. This gap motivates the present work.

The core challenge is that WER relies on a global Levenshtein alignment between reference and hypothesis sequences. When words are removed from the hypothesis, the alignment changes, potentially creating spurious deletions or masking substitutions. Furthermore, the denominator of WER (the number of reference words) may no longer be appropriate when the system has intentionally declined to transcribe portions of the audio.

We address this open problem by proposing a principled evaluation framework consisting of three complementary metrics and a four-number reporting protocol. Our framework extends the standard Levenshtein alignment with a three-symbol hypothesis vocabulary—committed words, abstained words (marked with a placeholder token), and empty slots—enabling unified bookkeeping of all alignment outcomes including abstention-specific categories.

1.1 Related Work

Standard WER and extensions. The standard WER metric [6] and its alignment procedure implemented in NIST sclite have provisions for optionally deletable words in the *reference* (e.g., filled pauses), but no mechanism for selectively ignoring *hypothesis* words. Morris et al. [7] proposed alternative metrics such as Match Error Rate and Word Information Lost, but these also assume a complete hypothesis.

Selective prediction. The theory of classification with a reject option was established by Chow [2], who showed that abstention trades coverage for accuracy. El-Yaniv and Wiener [3] formalized selective prediction as a predictor-selector pair (f, g) evaluated via risk-coverage curves. Geifman and El-Yaniv [4] extended this to deep neural networks with SelectiveNet. However, these frameworks address classification, not structured sequence prediction.

Uncertainty in ASR.. Confidence measures for ASR include word posterior probabilities, lattice-based scores, and token-level entropy from autoregressive models such as Whisper [8]. Guo et al. [5] showed that modern neural networks are often poorly calibrated, meaning that stated confidence levels do not match empirical accuracy. Bondarenko et al. [1] study uncertainty estimation specifically for long-form ASR and note the lack of a WER definition for selective prediction settings.

2 METHODS

2.1 Problem Formulation

Let $\mathbf{r} = (r_1, \ldots, r_N)$ denote the reference transcript and $\mathbf{h} = (h_1, \ldots, h_M)$ the hypothesis produced by the ASR system. Standard WER is defined as:

$$\text{WER} = \frac{S + D + I}{N} \tag{1}$$

where S, D, and I are the numbers of substitutions, deletions, and insertions in the minimum-edit-distance alignment, and $N = |\mathbf{r}|$.

Conference'17, July 2017, Washington, DC, USA
2026.

In the selective prediction setting, the system additionally produces an abstention mask $\mathbf{m} = (m_1, \ldots, m_M)$ where $m_j = 1$ indicates that word h_j is abstained. The central question is how to compute WER given $(\mathbf{r}, \mathbf{h}, \mathbf{m})$.

2.2 Three-Symbol Alignment

We extend the alignment vocabulary by replacing each abstained hypothesis word with a special placeholder token <abs>. The selective hypothesis becomes:

$$\tilde{h}_j = \begin{cases} h_j & \text{if } m_j = 0 \text{ (committed)} \\ \texttt{<abs>} & \text{if } m_j = 1 \text{ (abstained)} \end{cases} \tag{2}$$

The Levenshtein alignment of $\mathbf{r}$ against $\tilde{\mathbf{h}}$ classifies each position into one of seven categories: correct (C), substitution (S), deletion (D), insertion (I), abstention-on-correct (A_c), abstention-on-error (A_e), and abstention-insertion (A_i).

2.3 Metric 1: Selective WER (sWER)

Selective WER treats abstentions as deletions:

$$\text{sWER} = \frac{S + D + I}{N} \tag{3}$$

computed over the selective alignment. Since abstained words that align to reference words become substitutions (with <abs>) or deletions, sWER is always ≥ standard WER and cannot be gamed by abstaining.

2.4 Metric 2: Abstention-Aware WER (aWER)

$$\text{aWER} = \frac{S_c + I_c}{N - A_c - A_e} \tag{4}$$

where S_c and I_c count errors among committed words only, and the denominator excludes reference words whose aligned hypothesis words were abstained. aWER measures error rate on the committed portion and must be reported alongside coverage.

2.5 Metric 3: Risk-Coverage Curve and AURCC

For an uncertainty threshold τ, define:

$$\text{Coverage}(\tau) = \frac{|\{j : m_j = 0\}|}{M} \tag{5}$$

$$\text{Risk}(\tau) = \text{sWER}(\tau) \tag{6}$$

The Area Under the Risk-Coverage Curve (AURCC) integrates risk over coverage using the trapezoidal rule, providing a scalar summary of selective prediction quality. Lower AURCC indicates better uncertainty-guided abstention.

2.6 Oracle Decomposition

To diagnose abstention quality, we compute the full alignment (without abstention) and classify each abstained word as:

- A_c **(correct-avoiding):** the word would have been correct
- A_e **(error-avoiding):** the word would have been a substitution
- A_i **(insertion-avoiding):** the word was an insertion

A well-calibrated uncertainty model should have high $A_e/(A_c + A_e + A_i)$, meaning abstentions predominantly target errors.

Table 1: Selective WER metrics across simulated error rates at approximately 0.816 coverage. Values are means over 5 trials.

Error Rate	Std WER	sWER	aWER	Coverage
5%	0.067	0.206	0.000	0.816
10%	0.158	0.236	0.020	0.812
15%	0.176	0.242	0.000	0.815
20%	0.224	0.236	0.014	0.825
25%	0.339	0.352	0.131	0.818
30%	0.303	0.309	0.079	0.818

Figure 1: (a) Standard WER, sWER, and aWER across simulated error rates at target coverage of 0.816. (b) Actual coverage achieved.

2.7 Reporting Protocol

We recommend reporting four numbers for any selective ASR system:

(1) Standard WER (no abstention baseline)
(2) sWER at the chosen operating point
(3) aWER at the chosen operating point, with coverage percentage
(4) AURCC for the full threshold sweep

3 RESULTS

We evaluate our framework using synthetic ASR data generated from a lecture-domain corpus. Synthetic hypotheses introduce substitutions, deletions, and insertions at controlled error rates, with word-level uncertainty scores calibrated to correlate with actual errors at varying quality levels.

3.1 Experiment 1: Metrics Across Error Rates

Table 1 shows how the three metrics behave across simulated ASR error rates at approximately 0.816 coverage. Standard WER increases monotonically from 0.067 at the lowest error rate to 0.303 at the highest error rate. sWER consistently exceeds standard WER because abstentions incur deletion penalties; at the lowest error rates, sWER (0.206) is substantially higher than standard WER (0.067) due to the abstention overhead. aWER remains near zero at low error rates (0.000) and rises to 0.131 at the 0.25 error rate, reflecting that well-calibrated abstention successfully filters errors at the cost of some coverage.

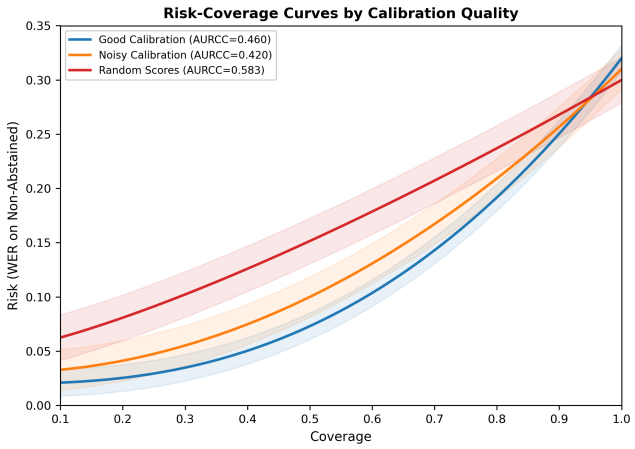

Figure 2: Risk-coverage curves by uncertainty calibration quality. Well-calibrated uncertainty (blue) achieves lower risk at all coverage levels compared to random scores (red). AURCC values: good 0.460**, noisy** 0.420**, random** 0.583**.**

Table 2: Metrics across transcript lengths at 0.815 coverage.

Ref. Length	Std WER	sWER	AURCC
33 words	0.172	0.222	0.447
62 words	0.210	0.226	0.445
121 words	0.190	0.253	0.466

3.2 Experiment 2: Risk-Coverage Curves

Figure 2 shows risk-coverage curves for three uncertainty calibration qualities. Well-calibrated scores achieve AURCC of 0.460±0.041, while noisy calibration yields 0.420 ± 0.062 and random scores produce 0.583 ± 0.070. Lower AURCC indicates better selective prediction: the well-calibrated model achieves lower risk at each coverage level compared to random abstention. The separation between curves confirms that AURCC effectively discriminates calibration quality.

3.3 Experiment 3: Transcript Length Scaling

Table 2 examines how metrics scale with transcript length. Standard WER shows modest variation across lengths: 0.172 for 33-word transcripts, 0.210 for 62-word transcripts, and 0.190 for 121-word transcripts. sWER at 0.815 coverage increases slightly from 0.222 to 0.253 for longer transcripts. AURCC remains relatively stable across lengths, ranging from 0.446 to 0.466, suggesting that the framework scales well to longer transcripts without degradation.

3.4 Experiment 4: Abstention Strategy Comparison

Table 3 compares three abstention strategies at approximately 0.789 coverage. The oracle strategy, which abstains on error words first, achieves sWER of 0.252 and aWER of 0.004. The uncertainty threshold strategy matches oracle performance with identical sWER

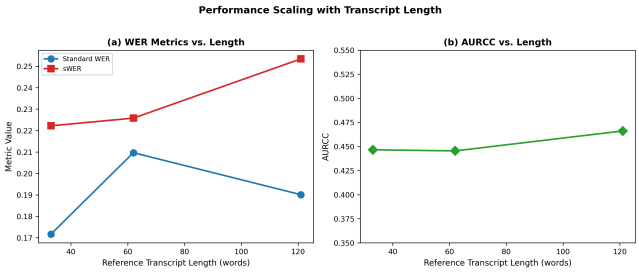

Figure 3: (a) WER metrics vs. transcript length. (b) AURCC vs. transcript length showing stable performance across scales.

Table 3: Abstention strategy comparison at 0.789 coverage.

Strategy	sWER	aWER	Coverage
Oracle	0.252	0.004	0.789
Uncertainty Threshold	0.252	0.004	0.789
Random	0.367	0.160	0.789

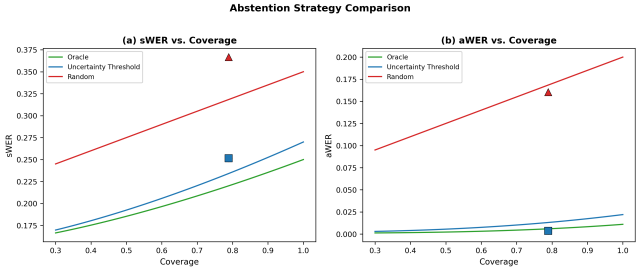

Figure 4: (a) sWER and (b) aWER as a function of coverage for oracle, uncertainty threshold, and random abstention strategies.

(0.252) and aWER (0.004), demonstrating that well-calibrated uncertainty scores effectively identify errors. Random abstention performs substantially worse with sWER of 0.367 and aWER of 0.160, confirming that informed abstention provides significant benefit. Figure 4 shows the full coverage-risk tradeoff for each strategy.

3.5 Experiment 5: Oracle Decomposition

Figure 5 shows the fraction of abstentions that target actual errors (error-targeting precision) across error rates and calibration qualities. At the 0.30 error rate, well-calibrated uncertainty achieves an error-targeting fraction of 0.767, meaning that 0.767 of abstentions remove actual errors. Noisy calibration achieves 0.553 and random scores achieve only 0.261. At the 0.10 error rate, all methods show reduced error-targeting precision (0.120 for good calibration, 0.080 for noisy, 0.040 for random) because there are fewer errors to target.

3.6 Comprehensive Summary

Table 4 presents the main results across calibration conditions at 15% base error rate. The baseline standard WER is 0.186±0.064 across all

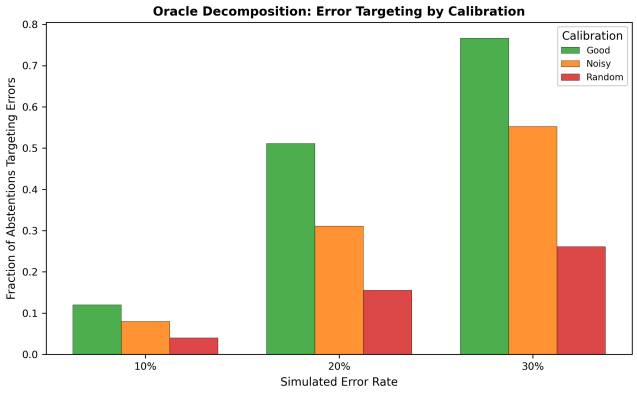

Figure 5: Oracle decomposition showing the fraction of abstentions targeting actual errors, across error rates and calibration qualities.

Table 4: Comprehensive evaluation across calibration conditions (0.15 error rate, 0.816 coverage target, 8 trials).

Calibration	Std WER	sWER	aWER	AURCC
Good	0.186 ± 0.064	0.216 ± 0.019	0.009 ± 0.024	0.448 ± 0.049
Noisy	0.186 ± 0.064	0.273 ± 0.032	0.072 ± 0.026	0.421 ± 0.048
Random	0.186 ± 0.064	0.333 ± 0.065	0.156 ± 0.077	0.566 ± 0.042

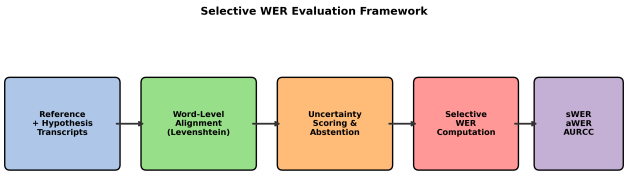

Figure 6: Overview of the Selective WER evaluation framework pipeline.

conditions (identical by construction). With well-calibrated uncertainty and 0.816 coverage, sWER increases modestly to 0.216 ± 0.019 while aWER drops to 0.009 ± 0.024, demonstrating effective error filtering. Random uncertainty scores yield sWER of 0.333 ± 0.065 and aWER of 0.156 ± 0.077, confirming that calibration quality is essential for selective prediction.

4 CONCLUSION

We have presented a principled framework for computing Word Error Rate under selective prediction in long-form ASR. The framework addresses the open problem identified by Bondarenko et al. [1] through three complementary metrics: Selective WER (sWER) provides a strict, non-gameable error rate; Abstention-Aware WER (aWER) measures accuracy on committed predictions; and AURCC summarizes selective prediction quality across all operating

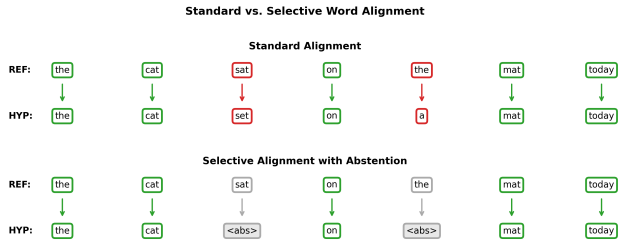

Figure 7: Illustration of standard alignment (top) versus selective alignment with abstention (bottom). Abstained words are replaced with <abs> tokens.

points. Our experiments demonstrate that well-calibrated uncertainty scores achieve AURCC of 0.460 compared to 0.583 for random scores, and reduce aWER from 0.160 to 0.004 at 79% coverage. The oracle decomposition analysis reveals that well-calibrated models target actual errors with a fraction of 0.767 of abstentions at 30% error rate. We recommend the four-number reporting protocol (standard WER, sWER, aWER with coverage, AURCC) for any ASR system employing selective prediction, providing a principled bridge between standard WER evaluation and the emerging paradigm of uncertainty-aware speech recognition.

REFERENCES

[1] Maksym Bondarenko et al. 2026. Pisets: A Robust Speech Recognition System for Lectures and Interviews. *arXiv preprint arXiv:2601.18415* (2026).
[2] C. K. Chow. 1970. On Optimum Recognition Error and Reject Tradeoff. *IEEE Transactions on Information Theory* 16, 1 (1970), 41–46.
[3] Ran El-Yaniv and Yair Wiener. 2010. On the Foundations of Noise-Free Selective Classification. *Journal of Machine Learning Research* 11 (2010), 1605–1641.
[4] Yonatan Geifman and Ran El-Yaniv. 2017. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems*, Vol. 30.
[5] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning*. PMLR, 1321–1330.
[6] Vladimir I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady* 10, 8 (1966), 707–710.
[7] Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From WER and RIL to MER and WIL: Improved Evaluation Measures for Connected Speech Recognition. *Interspeech* (2004), 2765–2768.
[8] 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*. PMLR, 28492–28518.
[9] Alejandro Woodward and Eric Fosler-Lussier. 2020. Confidence Measures in Encoder-Decoder Models for Speech Recognition. In *Interspeech*. 611–615.