# Exploring Lightweight Evaluation Metrics for Information Retention in ASR Outputs Beyond WER

*Ishana Rana*

[1]University of Zürich, Switzerland,

ishana.rana@uzh.ch

## Abstract

In the scope of my seminar in "Conversational Speech processing" at my university, I want to explore the evaluation metrics beyond the word rate error (WER) in automatic speech recognition (ASR) systems. WER has long served as the standard evaluation metric even though it fails to capture information retention and semantic similarity. This study explores a lightweight, semantically-informed evaluation approach, using metrics such as BLEU, ROUGE, BERTScore, and Named Entity Recognition (NER) overlap to assess ASR output quality. Using a subset of the AMI Meeting Corpus, transcriptions were generated with the Whisper model and evaluated under different conditions: with vs. without punctuation and scenario-based vs. natural meetings.

This work shows that simple, computationally efficient metrics can offer valuable complementary perspectives on ASR performance. Such approaches could be especially useful for academic coursework, low-resource development, and exploratory evaluation pipelines where interpretability and speed matter.

**Index Terms**: Automatic Speech Recognition, Evaluation Metric, Information Retention

## 1. Introduction

A common or even standard evaluation metric since quite the beginning of Automatic Speech Recognition (ASR) systems has been the Word Error Rate (WER) [1]. The WER is computed by aligning the predicted transcript with a reference and calculating the proportion of words that are substituted, deleted or inserted. Because of this, it serves to be a useful baseline metric for tasks where literal, word-for-word transcription accuracy is important - such as dictation or captioning systems. In such use cases, understanding the semantics or pragmatics of the content is not needed. The notable shortcomings of WER lie in conversational contexts, as it treats all errors equally. It ignores semantic similarity and fails to capture the impact of structural or linguistic variations, such as punctuations, word order or even grammar structures.

With the rise and advancement of ASR technologies in real-world applications, such as: voice assistants, automated meeting transcriptions, medical dictation or even human-machine interactions, the semantic coherence, or the usability of the ASR output often matters more than just the literal word-by-word accuracy. A high WER does not necessarily mean that the transcription is not useful in context, as it could still preserve key information that supports understanding or downstream processing. Consider the following example:

**Input audio:**
"We were just discussing of moving the meeting to friday."

**ASR system output:**
"We discuss moving meeting friday"

As one can see, the WER is high, but at the same time the information retention is still retained. This output, for example, could still be very useful and enough for a scheduling assistant to take the appropriate actions.

Motivated by this gap, this paper aims to explore a broader set of evaluation metrics that could better represent the communicative and semantic adequacy of ASR systems outputs - especially in the context of conversational speech. The focus lies on comparing ASR performance across two types of conversational data (natural and scenario-based), using a suite of lightweight but semantically informative metrics. These includes **BERTScore** (to assess semantic similarity), **ROUGE** and **BLEU** (for surface-level overlap) and **named entity recognition (NER) overlap** (to capture information content). **WER** is included as a baseline metric for comparison.

Lightweight metrics were chosen deliberately as they are faster to compute, require less memory and computational power, and are easy to implement in practical ASR development workflows. Which can be especially useful in low-resource settings, for prototyping, or for use cases where quick evaluation is needed—such as during iterative development, real-time system feedback, classroom projects or in explorative data analysis (EDA) settings. Despite their simplicity, these metrics can still capture insights on the amount of information retained in an ASR output.

## 2. Related Work

Because of the way WER fails to differentiate between content-preserving and meaning-altering substitution, there has been a growing amount of research investigating more nuanced evaluation methods with different approaches.

One style of approach for reference-based comparisons makes use of semantic similarity metrics. For example, *Kim et al. (2021)* propose the Semantic Distance metric, which uses sentence embeddings (like RoBERTa) to evaluate how much of meaning is lost in ASR output. The results of that paper show that the semantic metrics indeed correlate better with than WER on spoken language understanding (SLU) tasks [2]. Similarly, *Ngai and Rudzicz (2023)* also argue that not all recognition errors are equally bad for comprehension. They introduce severity measures that score ASR errors based on their impact on downstream sentiment and meaning.

Other works go beyond direct reference-based comparisons. Like the one of *Waheed et al. (2025)* propose a label-free evaluation model which approximates the ASR metrics using multimodal embeddings and learned proxies [3]. Furthermore,

there exist evaluation pipelines that incorporate large language models (LLMs), like the one of *Phukon et al. (2025)*. Their LLM-based metric is also combined with phonetic, semantic and inference-based components in order to approximate human judgements of speech clarity [4]. While both of these approaches show high correlation with human ratings and support real-time use, they are computationally quite expensive and require significant training or inference from large transformer models.

In contrast to the above, this paper, as already mentioned, aims to explore a set of lightweight yet semantically informative evaluation metric that are computationally efficient and easy to deploy. A recent related paper, that also focuses on the BLEU, ROUGE and BERTScore to evaluate Whisper-generated transcriptions of soccer game audio commentary, is the one from *Gautam et al. (2024)* [5]. Their work focuses on assessing how well ASR-generated transcripts support downstream tasks like event detection and game summarization. This paper also shares the interest in semantically-oriented ASR evaluation, but will differ in scope and aim: The focus here is specifically on conversational speech in a general-purpose setting (natural vs scenario based). Furthermore, the information retention itself is evaluated and not its effect on downstream model performance by comparing several types of metrics side-by-side.

# 3. Experimental Setup

## 3.1. Dataset Description

For the experiment in this paper, the Augmented Multi-party Interaction Meeting (AMI) Corpus [6] was used. It was obtained from the official AMI download site [7]. This dataset was chosen as it contains multi-speaker conversational recordings, clean audio channels, and detailed segment-level and word-level manual annotations. From the corpus, the following six meetings were selected: ES2016a–d (4) and EN2009c–d (2). The ES-meetings are scenario-based discussions (hence the "S"). This means participants follow a semi-scripted task (here: designing a remote control device). On the other hand the EN- meetings contain truly natural, unscripted speech (hence the "N"). The E in both cases refers to the recording location "Edinburgh". The four-digit number indicates the year and internal meeting index. The ES2016a–d subset comprises approximately 1h 56m 14s of audio, while EN2009c–d includes around 2h 33m 36s. Therefore, in total, roughly 4.5 hours of audio were analyzed. Both types of meetings were included to test for differences in ASR performance and evaluation metric behavior under more structured versus more spontaneous conversational conditions.

## 3.2. ASR System Configuration

All transcriptions were generated using the OpenAI Whisper model (base variant) [8], which was chosen because of its strong open-source performance in ASR tasks. To evaluate named entity recognition (NER), different model variants were tested: `dslim/bert-base-NER` [9], `Jean-Baptiste/roberta-large-ner-english` [10], and SpaCy's built-in NER pipeline [11]. These models were used to extract entities from only the ASR outputs.

## 3.3. Evaluation Metrics

**WER** was included as a baseline metric. In order to provide a more nuanced surface-level comparison than WER, **BLEU** and **ROUGE** were included to capture token-level and n-gram overlap. To asses semantic similarity at the sentence level using contextual embeddings, the **BERTScore** was analyzed. To quantify how well named entities are preserved in ASR outputs the **NER Overlap** was analyzed by evaluating **Precision**, **Recall** and the **F1-score**.

The gold-standard reference transcripts were constructed using the detailed manual annotations provided with the AMI dataset. It's important to note that named entity annotations were only available for the ES2016a–d meetings, which is why NER-based evaluation was conducted only on this subset.

## 3.4. Implementation and Tools

Whisper decoding was performed in Google Colab[12] to take advantage of faster GPU processing and memory. All other processing steps—NER tagging, metric computation, text normalization, and results analysis—were implemented in Python using a combination of HuggingFace Transformers, SpaCy, and custom scripts.

In addition to the Python-based processing, statistical analysis was conducted in R using RStudio. This included inspecting metric distributions and running statistical tests (such as independent and paired t-tests) to assess whether differences between meeting types or experimental conditions were statistically significant. The full codebase is available on my GitHub repository "Beyond-WER-in-ASR" [13].

# 4. Experiment

## 4.1. Dataset and Gold Standard Creation

Each meeting in the AMI dataset originally consists of multiple audio files, each corresponding to a different speaker channel. Since this experiment does not focus on speaker-specific variation, all channels per meeting were merged into a single file.

The corresponding gold-standard transcripts were generated by combining information from the `segments.xml` and `words.xml` annotation files provided in the AMI manual annotations. Similarly, for the NER evaluation, a gold-standard was generated by combining information from `words.xml`, `ne.xml` and `ontologies.xml`.

## 4.2. ASR Transcription with Whisper

All audio files were used as provided from the AMI corpus without additional pre-processing. The model outputs were in `.json` format. For further evaluation, the text from these files were extracted into `.txt` files.

## 4.3. Transcript Text Normalization Pipeline

In order to ensure a fair and consistent comparison between Whisper's outputs (hypotheses) and the gold-standard transcripts (references), a series of normalization steps were applied to both. These included: contraction expansion, disfluency removal, lowercasing, and extra space removal. Two versions were maintained throughout the experiments: one where punctuation was removed and one where it was preserved. This was done to evaluate whether punctuation had a measurable impact on evaluation metrics.

## 4.4. NER Extraction

To extract named entities from Whisper outputs, multiple methods were tested. It was clearly visible from the outputs that the models `dslim/bert-base-NER` and

`Jean-Baptiste/roberta-large-ner-english` performed poorly in terms of qualitative comparison with the gold-standard annotations.
Subsequently, SpaCy was used as an alternative, which provided better results overall. Two versions of SpaCy's pipeline were tested: `en_core_web_sm` (small model) and `en_core_web_trf` (transformer-based model). For each method, entities were extracted both with and without Inside–Outside–Beginning (IOB) tagging.

### 4.5. NER labels Normalization Pipeline

To fairly compare SpaCy extracted entities against the gold-standard, a normalization step was applied to the entity labels. A Python script was written to first visually see the differences between the gold and predicted NER outputs. Due to misalignment issues caused by IOB tagging and token mismatch, only the plain label format (without IOB tags) was used for the final evaluation.
Furthermore, since the original AMI NER tags included domain-specific labels (e.g., "PROJECT MANAGER") that are not present in the SpaCy label set, all gold-standard tags were mapped to SpaCy-compatible categories. This ensured a fair comparison and minimized label mismatch.

### 4.6. Computing Evaluation Metrics

All evaluation metrics were computed using custom Python scripts, available on the GitHub repository. Outputs were saved in `.csv` format to facilitate easier statistical testing in R and to allow for direct visual inspection.

### 4.7. Analysis in R

#### *4.7.1. Transcription-based Evaluation*

Two main conditions were tested:
**(1)** the presence or absence of punctuation in the transcripts
**(2)** the meeting type—scenario-based (ES) versus natural (EN).
The evaluation scores for each metric (WER, BLEU, ROUGE, BERTScore) were visualized using bar plots and line plots to assess their distributions across conditions. Prior to conducting statistical tests, normality assumptions were checked with the Shapiro–Wilk test and by visualizing the QQ-plots. All distributions passed the normality checks.
A paired t-test was used to compare the influence of punctuation on the evaluation scores. Independent two-sample t-test were used to compare the scores between scenario-based and natural meetings also considering the punctuation condition. For each case, a 95% confidence interval was computed to test the following hypotheses:

- Whether there is a statistically significant difference in evaluation scores with vs. without punctuation.
- Whether there is a statistically significant difference between scores of natural and scenario-based meetings.

#### *4.7.2. NER-based Evaluation*

For named entity recognition, the performance of the two SpaCy models (`spacy_sm` and `spacy_trf`) was evaluated using precision, recall, and F1-score. These scores were visualized using bar plots for both models.

## 5. Results

### 5.1. Evaluation Metric Score

Table 1 shows the average evaluation metric scores for all meetings under the two transcription conditions: with punctuation and without punctuation. The metrics include WER, BLEU, ROUGE-1, ROUGE-L, and BERTScore. The exact scores, processing scripts, and plots for all evaluation conditions can be found in the GitHub repository [13].

| Metric | With Punctuation | | Without Punctuation | |
|---|---|---|---|---|
| | ES- | EN- | ES- | EN- |
| WER | 87.67 | 84.61 | 81.99 | 77.41 |
| BLEU | 43.39 | 46.29 | 40.31 | 43.46 |
| ROUGE-1 | 78.24 | 78.18 | 76.52 | 76.36 |
| ROUGE-L | 37.77 | 37.71 | 39.24 | 39.07 |
| BERTScore | 86.94 | 88.77 | 91.82 | 92.98 |

Table 1: *Table of average metric scores by meeting type and punctuation condition.*

The same scores can also be visually inspected in Figure 1 below, which presents the same data but as grouped bar plots.

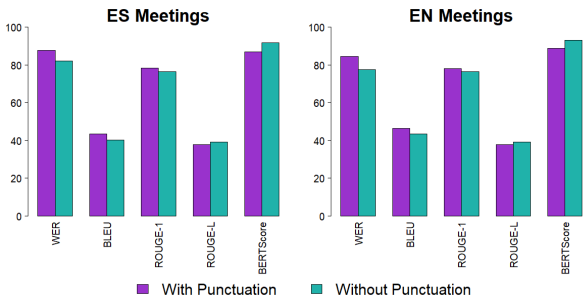


Figure 1: *Barplots of average metric scores by meeting type and punctuation condition*

As we can see, the punctuation condition had the greatest effect on the WER metric, where transcripts with punctuation consistently scored worse than those without. BLEU had the second largest average difference. It had slightly higher scores when punctuation was included. In contrast, ROUGE-L was least affected by punctuation presence.
As a sidenote: for WER, lower scores indicate better performance, while for BLEU, ROUGE, and BERTScore, higher values are better. The results show that BERTScore consistently rated the quality of the ASR outputs higher than the surface-level metrics (WER, BLEU, ROUGE).
The distinction between ROUGE-1 and ROUGE-L lies in what kind of overlap they measure. ROUGE-1 considers single-word overlap, while ROUGE-L captures the longest common subsequence (LCS), which is more reflective of in-order matching and structural similarity.
Overall, punctuation has the strongest negative impact on WER, modestly increases BLEU and ROUGE-1 scores, has minimal influence on ROUGE-L, and appears to slightly improve BERTScore when removed.

#### 5.1.1. Statistical Test Results

To assess whether the observed differences in metric scores between punctuation conditions and meeting types were statistically significant, normality tests were first conducted as a prerequisite for applying parametric tests. It is important to note that the sample sizes for these tests were relatively small, and this should be kept in mind when interpreting the results.
The Shapiro–Wilk test was applied to each metric under both conditions—punctuation (with vs. without) and meeting type (ES vs. EN)—to test the assumption of normality. In addition, QQ-plots were used to visually assess the score distributions. For all metrics, the results indicated that the assumption of normality was reasonable, as all p-values were bigger than 0.05. Therefore, I failed to reject the Null hypothesis, that the data comes from a normally distributed population. Table 2 summarizes the p-values of the Shapiro–Wilk tests for each metric under both comparison conditions.

| **Metric** | **Punctuation Cond.** | **Meeting Type Cond.** | |
|---|---|---|---|
| | | **punct.** | **no punct.** |
| WER | 0.477 | 0.635 | 0.830 |
| BLEU | 0.150 | 0.657 | 0.452 |
| ROUGE-1 | 0.331 | 0.312 | 0.322 |
| ROUGE-L | 0.103 | 0.920 | 0.923 |
| BERTScore | 0.937 | 0.748 | 0.833 |

Table 2: *P-values of Shapiro–Wilk normality tests for each metric under punctuation and meeting-type conditions.*

As normality may be assumed, paired t-tests were conducted to evaluate the difference in scores between the punctuation conditions. Two-sample t-tests were used to assess the differences in metric scores between the two meeting types also considering each punctuation condition separately. Table 3 summarizes the resulting p-values from these tests. Scores in bold indicate statistically significant differences at the $\alpha = 0.05$ level. Among all comparisons, only the punctuation condition (with vs. without) showed statistically significant differences across almost all five metrics (exception was ROUGE-L). The meeting types did not yield significant differences for any metric under either punctuation condition.

| **Metric** | **Paired t-test** | **2 sample t-test** | |
|---|---|---|---|
| | | **punct.** | **no punct.** |
| WER | **0.0009** | 0.1173 | 0.1169 |
| BLEU | **0.0375** | 0.1173 | 0.1169 |
| ROUGE-1 | **0.0433** | 0.1908 | 0.1950 |
| ROUGE-L | 0.0986 | 0.7052 | 0.7146 |
| BERTScore | **0.0032** | 0.2824 | 0.2458 |

Table 3: *P-values from paired and two-sample t-tests for each evaluation metric. Statistically significant results ($p < 0.05$) are in bold.*

#### 5.1.2. Multivariate Analysis

Additionally, a multivariate analysis was conducted to explore potential relationships between the evaluation metrics with the help of correlation matrices. Figure 2 presents the resulting correlation matrices considering the punctuation condition.

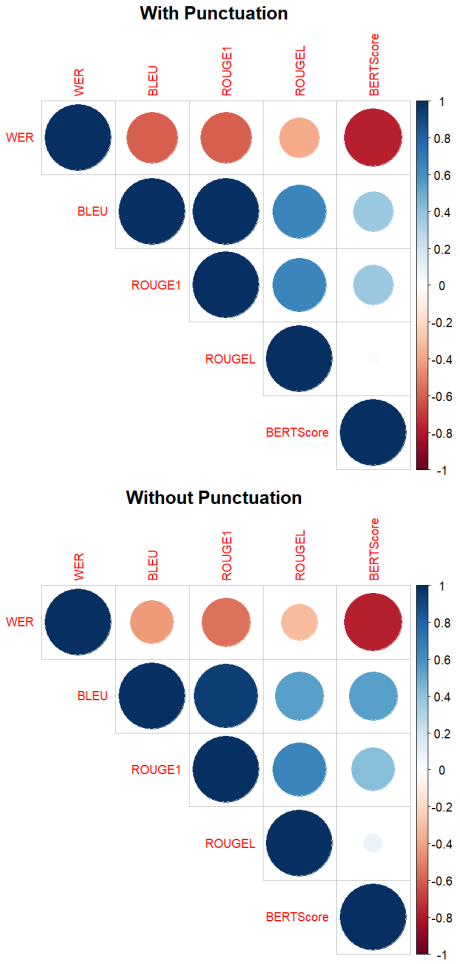


Figure 2: *Spearman correlation matrices between evaluation metrics with punctuation and without punctuation, highlighting how punctuation influences inter-metric relationships.*

Follow-up statistical tests were performed on some selected pairs to test whether the observed correlations were statistically significant at the $\alpha = 0.05$ level. The results of these tests are summarized in Table 4. The results where punctuation was not included, is not only available on GitHub, as none of the pairs showed any significant correlation.

### 5.2. NER Evaluation Results

In order to evaluate how well named entities were preserved in the ASR output, the following scores were computed: exact, partial, missed, spurious, precision, recall and F1 scores. The primary difference between the two spaCy models lies in their underlying architectures—SM is a lightweight statistical model, whereas TRF leverages transformer-based contextual embeddings for more robust and accurate predictions. The averaged results across all ES meetings are presented in Table 5. Full per-file results can be found in the linked GitHub repository.

| Metric Pair | roh | p-val. | Interpretation |
|---|---|---|---|
| WER-BERTScore | -0.77 | 0.10 | neg. trend, but not significant |
| WER-ROUGE-1 | -0.60 | 0.24 | moderate inverse trend |
| WER-ROUGE-L | -0.37 | 0.50 | Weak inversive trend |
| **BLEU-ROUGE-1** | **1.00** | **0.0028** | **pos. correlation** |

Table 4: *Spearman correlation matrix plot for the condition with punctuation. Statistically significant correlations ($p < 0.05$) are highlighted in bold.*

| Metric | SpaCy SM | SpaCy TRF |
|---|---|---|
| Exact | 6.75 | 7.00 |
| Partial | 3.25 | 5.50 |
| Missed | 62.25 | 59.75 |
| Spurious | 20.00 | 17.00 |
| Precision | 0.21 | 0.27 |
| Recall | 0.08 | 0.10 |
| F1 Score | 0.11 | 0.14 |

Table 5: *Averaged NER scores across ES meetings for SpaCy's small (SM) and transformer-based (TRF) models.*

# 6. Discussion and Analysis

## 6.1. Metric Behavior Across Conditions

As shown by the statistical tests, the inclusion or exclusion of punctuation in the evaluation pipeline had a significant impact on the metric scores. Both conditions were deliberately tested, as punctuation plays an important role in many downstream applications. In tasks such as summarization, dialogue modeling, or intent detection, punctuation not only aids in readability but also helps disambiguate meaning and set out sentence boundaries—therefore, contributing to better semantic understanding. The metrics ROUGE-1 and ROUGE-L were included specifically to capture both content overlap and structural accuracy. ROUGE-1 gives insight into lexical preservation, while ROUGE-L is more tolerant to minor disfluencies and reordering, which is especially suitable for conversational speech. This behavior is also reflected in the results, where the punctuation condition had the least influence on ROUGE-L, showing that it is more robust to superficial variations in structure.

As we can see from the results, the metrics BLEU and ROUGE-1 performed better when punctuation was included for both meeting types. This suggests that these surface-level metrics benefit from punctuation cues, likely because punctuation improves token boundary alignment and increases n-gram match rates. In contrast, metrics like BERTScore and especially WER performed slightly worse when punctuation was included. WER likely suffered more because it relies strictly on word-to-word matching, and punctuation can introduce alignment mismatches that inflate error counts. The drop for BERTScore seemed counterintuitive at first, since it is a semantic metric. However, it is possible that the inclusion of punctuation introduced embeddings that slightly misaligned more with the semantic representation. Still, the overall difference is still small. The smaller performance changes in BLEU, ROUGE, and BERTScore suggest that these metrics are more robust and better able to capture semantic similarity or contextual preservation beyond literal matches.

## 6.2. Meeting Type Comparison: Scenario vs. Natural

The decision to include both scenario-based and natural meeting types was made to examine ASR evaluation performance under two distinct forms of conversational speech. The scenario-based meetings (ES series) followed a semi-scripted structure, in which participants were guided by a predefined task. In contrast, the natural meetings (EN series) featured fully unscripted, spontaneous conversations. These were intended to better simulate real-world usage conditions for ASR systems, where speech is less structured, more prone to hesitations, overlapping talk, and interruptions. Including both types allowed for a more realistic assessment of metric behavior across different conversational contexts—structured versus spontaneous—thus supporting the goal of exploring semantic evaluation in practical applications.

As such, I had expected the ASR performance to be higher in the scripted meetings due to reduced noise and less or no speech overlap in their meeting audio. The results from the statistical tests showed no significant differences in metric scores between the two meeting types under either punctuation condition. While this might suggest that ASR performance and metric behavior are comparable across structured and unstructured speech, it is important to note that the statistical power of these tests is limited.

## 6.3. Inter-metric Relationships

To better understand how the different evaluation metrics relate to one another and whether they capture overlapping or complementary information, a correlation analysis was conducted. This was particularly relevant given the hybrid nature of the selected metrics—ranging from surface-level string comparisons (e.g., WER, BLEU, ROUGE) to deeper, embedding-based semantic measures like BERTScore.

As the results in Table 4 show, Word Error Rate (WER) exhibited a consistently negative correlation with all other metrics. This is expected, as WER is an error-based metric where lower values indicate better performance, while the other metrics increase with improved output quality. The strongest negative correlation was observed between WER and BERTScore ($\rho \approx -0.77$, p = .10). Although not statistically significant at the $\alpha$ = 0.05 level, this suggests a clear trend: transcripts with more literal word errors (high WER) also tend to diverge semantically from the reference (lower BERTScore).

The only statistically significant result was the perfect positive correlation between BLEU and ROUGE-1 ($\rho$= 1.00, p = .0028), indicating that these two surface-based metrics are highly aligned in how they assess lexical overlap. This could be due to the fact that both rely on n-gram comparisons—BLEU on precision, and ROUGE-1 on recall—making them especially sensitive to similar word-level content matches.

In contrast, ROUGE-L showed minimal correlation with BERTScore ($\rho \approx 0$, p = 1.0), suggesting that these two metrics are measuring entirely different aspects of transcript quality. This weak association highlights that a high structural match (e.g., in sentence order or flow) does not necessarily imply high semantic preservation, and vice versa.

### 6.4. NER Findings in Context

Good NER overlap can be very helpful in downstream tasks such as meeting summarization, dialogue retrieval, and question answering—where key roles, locations, or project names often carry the most crucial information.
The results of the NER-based evaluation showed overall low scores for both SpaCy models. One major reason may lie in the nature of the AMI annotations themselves: the named entity tags are highly domain-specific and fine-grained (e.g., "PROJECT MANAGER", "DEVICE"), which are not part of the standard label sets used by general-purpose NER systems. Even after harmonizing the gold-standard annotations to match SpaCy's tag schema, the models failed to consistently detect relevant entities. Many named entities were missed entirely, which negatively impacted recall.
This mismatch indicates that the poor NER performance is likely not due to the ASR output quality itself, but rather due to limitations in the NER models' ability to generalize to the specialized context of AMI meetings. Interestingly, this is also reflected in the fact that all other evaluation metrics (BLEU, ROUGE, BERTScore) consistently rated the ASR output quality higher than the NER-based analysis did.
As expected, the transformer-based SpaCy model (`en_core_web_trf`) outperformed the smaller statistical model (`en_core_web_sm`) in terms of precision, recall, and F1 score. This is likely due to its use of contextualized embeddings, which allow it to better capture the meaning and function of entities in sentence-level context.

## 7. Limitations and Future Work

While this study offers a first exploratory look into lightweight, semantically informed evaluation of ASR outputs in conversational contexts, there are several limitations that should be acknowledged. Most notably, the statistical power of the tests conducted in this study is limited. Although the total audio analyzed spans about 4.5 hours, the number of meetings included is relatively small. This reduces the ability of the statistical tests to detect more subtle but potentially meaningful differences. Therefore, the findings should be interpreted with caution and seen as indicative rather than conclusive.
A more comprehensive evaluation would ideally include the full AMI corpus, which consists of approximately 100 hours of meeting recordings. It improve the robustness of statistical analyses, allow for better generalization across conditions, and make it possible to conduct more in-depth analyses (e.g., turn-level, speaker-role-based, or topic-specific evaluations). However, working at this scale would require significantly more computational resources and processing time—particularly for decoding, entity tagging, and metric computation. In addition to scaling up, future work could explore more advanced semantic evaluation methods. Such as meaning-based edit distance, semantic distance with sentence embeddings, or LLM-based scoring.

## 8. Conclusions

This study explored whether lightweight, semantically informed metrics offer a more nuanced view of ASR quality than WER alone.
The results show that alternative metrics can provide complementary insights: BLEU and ROUGE-1 aligned closely, while BERTScore offered deeper semantic assessment. WER, in contrast, remained sensitive to token-level mismatches—especially with punctuation. Notably, transcripts with high WER still scored well on the other scores, highlighting the risks of relying solely on WER.
Punctuation had a significant effect on most metrics, particularly WER, while meeting type (scenario-based vs. natural) showed no statistically significant differences—though limited sample size restricts strong conclusions.
NER-based evaluation revealed limitations of general-purpose models, which failed to capture many domain-specific entities, leading to low recall and F1 scores.
Overall, the findings suggest that even simple, efficient metrics can capture meaningful differences in ASR output. For academic coursework, low-resource prototyping, and exploratory analysis, such metrics could make a practical step towards a more context-aware ASR evaluation that goes *beyond WER*.

The Interspeech 2025 organisers would like to thank ISCA and the organising committees of past Interspeech conferences for their help and for kindly providing the previous version of this template.

## 9. References

[1] S. Goldwater, D. Jurafsky, and C. D. Manning, "Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, no. 3, pp. 181–200, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639309001599

[2] H. Kim *et al.*, "Semantic distance: A new metric for asr performance analysis towards spoken language understanding," in *Proceedings of Interspeech 2021*, 2021.

[3] A. Waheed *et al.*, "On the robust approximation of asr metrics," *arXiv preprint arXiv:2502.12408*, 2025.

[4] D. Phukon *et al.*, "Aligning asr evaluation with human and llm judgments: Intelligibility metrics using phonetic, semantic, and nli approaches," *arXiv preprint arXiv:2506.16528*, 2025.

[5] S. Gautam, M. H. Sarkhoosh, J. Held, C. Midoglu, A. Cioppa, S. Giancola, V. Thambawita, M. A. Riegler, P. Halvorsen, and M. Shah, "Soccernet-echoes: A soccer game audio commentary dataset," in *2024 International Symposium on Multimedia (ISM)*, 2024, pp. 71–78.

[6] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The ami meeting corpus: A pre-announcement," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2005, pp. 24–35.

[7] AMI Consortium, "Ami meeting corpus - download page," https://groups.inf.ed.ac.uk/ami/download/, accessed: 7th July 2025.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Whisper: Robust speech recognition via large-scale weak supervision," https://github.com/openai/whisper, 2022, accessed: 2025-07-15.

[9] D. Davlan, A. Ozgur, and N. I. Cetin, "Bert-ner: Fine-tuned bert model for named entity recognition," https://huggingface.co/dslim/bert-base-NER, 2021, accessed: 2025-07-15.

[10] Jean-Baptiste, "roberta-large-ner-english," https://huggingface.co/Jean-Baptiste/roberta-large-ner-english, 2022, accessed: 2025-07-15.

[11] M. Honnibal, I. Montani, S. V. Landeghem, and A. Boyd, "spacy: Industrial-strength natural language processing in python," https://spacy.io, 2020, accessed: 2025-07-15.

[12] I. Rana, "Beyond-wer-in-asr – google colab notebook," https://colab.research.google.com/drive/1Rx3XCCquKNVfCN2B3PwPXMy5BT56dm0M?usp=sharing, 2025.

[13] Juiceifers, "Beyond-wer-in-asr," 2025, gitHub repository. [Online]. Available: https://github.com/Juiceifers/Beyond-WER-in-ASR