Episode Book

EPISODE #5962

How Word Error Rate Lies to Us

PUBLISHED	14 October 2026 at 14:40 UTC
DURATION	26:49
LISTEN	myweirdprompts.com/episode/word-error-rate-limits-05326/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	deepseek-flash
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.4
SOURCES	21 cited by the research (listed at the end)
BOOK GENERATED	19 October 2026 at 17:42 UTC

SYNOPSIS

Word error rate has a formula that fits on a napkin, yet it's the number that ends up in every headline about speech recognition. This episode digs into the mechanics: how edit distance at the word level gets computed, why WER has no upper bound, and when character error rate is the more honest lens. Then it goes after everything a single score buries — normalization choices that swing results by forty points, a Whisper normalizer that strips vowel signs out of Indic scripts, inverse text normalization for numbers, named entities, code-switching, punctuation, and speaker attribution. Along the way: why WER agrees with human judgment less than any other metric tested, and what semantic metrics are doing about it.

TRANSCRIPT



Corn

Here's the question that's been rattling around my head since Daniel sent this in. How does a transcript with a ninety-five percent on the spec sheet end up being the one nobody can use?



Herman

And it's a fair question, because the metric behind that ninety-five percent has a formula that fits on a napkin.



Corn

Daniel's prompting us on the accuracy half of ASR benchmarking this time. Real-time factor, latency, throughput, what it costs to run the thing. Today it's the other half. How do we actually turn a transcript into a number, what does that number measure, and where does it lie to us. He wants the mechanics of word error rate, the case for character error rate, all the stuff a single score buries. Punctuation, casing, numbers, proper nouns, accents, code-switching, normalization. The qualitative side too. Readability, hallucinations, semantic fidelity, speaker attribution. And then how the benchmarks themselves get built and whether they hold up.



Herman

Which is really two questions wearing one coat. How do you compute the score, and what does the score actually mean.



Corn

And one of those has a clean answer.



Herman

Right, so let's start with the metric everyone reports, and the formula underneath it. Because the formula is simple. The interesting part is what happens when you apply it.



Herman

Word error rate. WER. It's an edit distance, computed at the word level rather than the phoneme level. The formula is substitutions plus deletions plus insertions, all over N, where N is the number of words in the reference transcript. And N itself breaks down as substitutions plus deletions plus correct words.



Corn

So it's the Levenshtein distance, just measured on words instead of characters.



Herman

Exactly that. It's asking a mechanical question. How many edits does it take to turn the model's output back into the reference. Three kinds of edits count. A substitution is one word swapped for another. A deletion is a word in the reference that the model didn't produce. An insertion is a word the model produced that wasn't in the reference.



Corn

Deletion versus insertion trips people up. Whether you say the reference has a missing word or the hypothesis has an extra one depends on which way you're walking.



Herman

It's the direction of travel that decides it. The reference is the truth, so if the model drops a word, that's a deletion. One way to hold onto it is to ask, how do I get from the reference to the hypothesis. If the reference says 'this is wikipedia' and the hypothesis says 'this wikipedia', you're getting there by deleting a word. That's a deletion.



Corn

Give me the worked example. Small one.



Herman

Reference is 'the cat sat on the mat'. Six words. The model outputs 'the cat sit on the'. Same first four words, roughly, but 'sit' where the reference has 'sat', and then it just... stops. So the alignment gives you one substitution, 'sit' for 'sat'. Zero insertions. One deletion, because 'mat' never shows up.



Corn

And the denominator is the reference length, so six.



Herman

Two errors over six reference words. WER is zero point three three three. Thirty-three point three percent.



Corn

All that for a cat.



Herman

And the alignment isn't done by eye. It's a dynamic programming problem. You build an edit distance matrix and walk the cheapest path through it. That's what jiwer does, what Hugging Face evaluate does, SpeechBrain's edit distance module, TorchMetrics, Lingvo's simple wer. Same machinery under all of them.



Corn

Here's the part I want to poke at. When you run this over a whole test set, do you average the per sentence scores?



Herman

No, and this matters more than it sounds. You aggregate S, D, I and N across every sentence first, and then you divide once at the end. You do not average the per sentence rates.



Corn

Why not.



Herman

Because a short sentence carries the same weight as a long one if you average the rates. Say a model mangles one three-word utterance and nails a forty-word one. Per-sentence averaging treats those as equal events. Aggregating the edit counts weights them by how much speech was actually in them. The aggregated number is the one that reflects the corpus. The averaged one usually flatters a model that's bad on short utterances, and short utterances are most of what a dictation tool sees.



Corn

So the number everybody quotes has a weighting choice baked into it that most people never look at.



Herman

And it's the right choice, but you have to know it's there.



Corn

You said WER has no upper bound. Walk me through that, because it sounds like a typo.



Herman

It's real. Because N is the reference length, and insertions live in the numerator. If a model produces far more words than the reference, the numerator can grow past the denominator. The number goes above one. Hugging Face's example is a two-word reference where the model outputs ten words. That gives you a WER of five. Five hundred percent.



Corn

A model that talks too much can score five hundred percent. My kind of failure.



Herman

And the companion metric, word accuracy, is one minus WER, which they also write as correct words minus insertions over N. So when WER goes above one, word accuracy goes negative. You can have an accuracy of minus two hundred percent.



Corn

A percentage of what, exactly.



Herman

Of nothing. It's an accounting identity that stopped being a percentage somewhere around one.



Corn

Same example at the character level, though. That's where it gets interesting.



Herman

Same pair. 'The cat sat on the mat' against 'the cat sit on the'. One substitution at character level, the t in sit, zero insertions, four deletions for the missing letters in 'mat' and the space and the final word. Twenty-two characters in the reference. CER comes out to zero point two two seven. Twenty-two point seven percent.



Corn

Thirty-three at word level, twenty-three at character level. One character off in 'sit' and the whole word gets marked wrong.



Herman

That's the thing the Hugging Face course calls out directly. WER penalizes spelling errors heavily no matter how minor. 'Sit' is a single character away from 'sat', and at the word level that's a full error. At the character level you get partial credit for being close.



Corn

So when is CER the right lens.



Herman

Whenever the notion of a word boundary is doing less work than the notion of a character. Languages written without spaces, Japanese, Thai, Chinese. There's no word to count. A word level score has no boundaries to find. CER is the honest metric there because it's measuring what's actually on the page.



Corn

That's not just a non-Latin script problem, though. There's a 2025 paper, Thennal and co-authors, arguing CER correlates more closely with human judgment than WER even for English. Their claim is WER fails in specific ways with morphologically complex languages and languages without clear word boundaries, and they ran human evaluations in Malayalam, English and Arabic to check it.



Herman

And it lines up with the newer work from this year. Sharma and co-authors went looking at which metric agrees with people, and WER came in last. Least agreement with human judgment of everything they tested. The best semantic metrics scored highest, and their line on CER is that despite its simplicity and low cost it stays remarkably close to those best configurations. They recommend CER plus a semantic distance metric as complementary.



Corn

The metric everybody reports is the one that agrees with humans the least.



Herman

Sixty-three percent agreement for WER, against ninety-two to ninety-four percent for the best LLM-based hypothesis selection. That's the gap. The thing on the leaderboard is not the thing a person would pick.



Corn

Japanese is the case that makes it vivid for me. Most words there can be spelled more than one way. A system transcribes 'adviser' where the ground truth says 'advisor', and that counts as an error. Same word.



Herman

Same word, different spelling convention, full substitution. The lenient evaluation work on Japanese speech recognition is exactly about that. And the scoring doesn't care.



Corn

Here's what I want to hold onto. The formula measures edit distance. It does not measure meaning. It has never claimed to. But it's the number everybody puts in the headline.



Herman

And it's cheap, reproducible, comparable across systems. That combination is hard to beat even when the metric is wrong. You can run WER on a laptop. You can compare two systems that were built years apart. The number is stable. Meaning is expensive to measure.



Corn

So WER and CER give us numbers. The question is what those numbers hide. And that's where things get messy.



Herman

Start with normalization, because it's the biggest hidden lever and it's the least visible from outside.



Corn

Define it for me. What is being normalized.



Herman

Lowercasing. Stripping punctuation. Sometimes collapsing whitespace, sometimes expanding or contracting numbers. The idea is to compare transcripts on content rather than on formatting, so a model doesn't get punished for choosing a comma where the reference didn't. The Whisper paper itself flags the drastic effect normalization has on WER.



Corn

How drastic.



Herman

Take Whisper small on Dhivehi from Common Voice. Orthographic WER is a hundred and sixty-seven percent. Run it through the normalizer and it drops to a hundred and twenty-six. Forty-one points of swing from a text preprocessing step.



Corn

That's not a rounding difference, that's the difference between a system that looks broken and one that looks merely bad. And it's a choice somebody made in a config file.



Herman

Which would be fine if the normalizer were neutral. The problem is it isn't. Whisper ships a non-English normalizer that is actively harmful for Indic scripts. Manohar and co-authors went through it character by character. It strips Unicode mark class characters. Those include vowel signs, the matras, and the virama. Strip those and you're left with bare consonants. The words fall apart.



Corn

Give me the measure of how much they fall apart.



Herman

They used METEOR similarity to compare text before and after normalization. English is zero point nine seven. Finnish zero point nine five. So for those, normalization barely changes anything, which is the intent. Hindi comes in at zero point three eight. Malayalam and Tamil are essentially zero. Thai is zero flat. The text after normalization has nothing in common with the text before it.



Corn

And the WER numbers that come out of that?



Herman

Absolute reductions on Whisper small over FLEURS. Twenty-one point nine percent for Hindi. Forty-one point five for Tamil. A hundred and fifty-two point two percent for Malayalam. The authors' conclusion is that current normalization practices artificially boost WER and SOTA performance, and they call for linguistically informed normalizers built by native speakers.



Corn

Nobody shipped a fix.



Herman

This routine went into Meta's MMS, Seamless, AssemblyAI's Conformer-1, the Hugging Face transformers library. It's everywhere. And it's still producing numbers that flatter systems on exactly the languages where the evaluation is hardest.



Corn

So the leaderboard for Malayalam is partly a measurement of how aggressively the normalizer deleted the language.



Herman

That's a fair way to put it. And there's a second text problem sitting next to it, which is inverse text normalization. Going from spoken form to written form. Turning 'twenty five dollars' into the numerals and the dollar sign. NeMo has a weighted finite state transducer library for it, and their own documentation says the task has extremely low tolerance for unrecoverable errors. Because if you get it wrong, you've changed a number, and a changed number in a medical note is not a typo.



Corn

Named entities and numbers. Altinok's work flags Whisper specifically as struggling with named entities and numerical data, especially when formatting is required. And that shows up hardest in legal, financial and medical domains.



Herman

Where the number is the content.



Corn

So the model gets the sentence right and the account balance wrong.



Herman

And here's the question I find interesting. Can two transcripts have meaningfully different WER scores and still be equally useful to a reader?



Corn

I'd bet yes.



Herman

McNamara and co-authors make the case that existing WER reports significantly overestimate the number of contentful errors in state of the art systems. Their method is multiple references. Instead of one ground truth transcript, they transcribe the same audio under opposing style parameters. Different but valid ways of writing down the same speech. Score against a single reference and you're penalizing style. Score against several and you can separate a real error from a legitimate choice.



Corn

The Swiss German study is the one that makes it concrete for me. Numbers.



Herman

Measured WER twenty-five point six percent. They then separate genuine errors from valid stylistic variation. Tense, word order, Swiss orthography. Content WER drops to thirteen point eight. Then they correct for estimation bias and it lands at eight point five. The authors' conclusion is that the true error rate is roughly a third of measured WER.



Corn

Two thirds of the reported errors were the model writing Swiss German differently and not wrongly.



Herman

And that number, twenty-five point six, is the one that gets quoted.



Corn

Now take the case where the transcript looks clean. No errors at all. Is that safe.



Herman

This is where hallucinations come in, and it's the part I'd want every clinician to hear. Hallucinations are fluent, coherent transcript output that has nothing to do with the audio. They're worse than an obvious failure because they preserve the grammatical and



Herman (continued)

semantic structure. They read like a real transcript. The SHALLOW benchmark categorizes them along four axes. Lexical, phonetic, morphological, semantic.



Corn

And they happen on nearly perfect transcripts.



Herman

That's the finding from HALAS, the first human annotated dataset of naturally occurring hallucinations. Seven state of the art ASR models, recorded on real earnings calls. Their line is that hallucinations also occur for almost correctly transcribed speech, the kind characterized by a low word error rate.



Corn

So a system can be at two percent WER and invent a sentence in the middle. And the metric that everyone uses would call that an excellent transcript.



Herman

Their detection numbers are the other half of it. Character and semantic metrics used as hallucination proxies reach eighty-one percent ROC AUC. The current state of the art detectors get an F one of fifty-three point one percent. Barely better than a coin flip for finding them.



Corn

People are deploying systems that are excellent at the metric and undetectable at the failure.



Herman

There is mitigation work. One paper projects decoder activations down to a low rank subspace, and it cuts the average hallucination rate from thirty-one point three one percent down to two point four four. That's a ninety-two percent relative reduction on



Herman (continued)

non-speech benchmarks. The cost is zero point three three to four point three nine percentage points of WER on LibriSpeech.



Corn

They traded a small WER hit for a large hallucination reduction. Which tells you what they think the more dangerous number is.



Herman

And it's the right trade in a medical context, where a hallucinated sentence in a patient's chart is worse than a missing word.



Corn

So the field knows. It's building metrics to catch what WER misses.



Herman

It's building a lot of them. EmbER is embedding error rate, weighting errors by semantic distance. POSER is part of speech error rate. S squared ER is a sentence level semantic error rate that uses an LLM as the judge. There's an intelligibility metric combining natural language inference scores, semantic similarity and phonetic similarity that reaches zero point eight nine correlation with human judgments on the Speech Accessibility Project data for dysarthric and dysphonic speech. That's a domain where a high WER and clear human understanding sit side by side, because the meaning survives even when the words don't.



Corn

Does one of these win.



Herman

No. There's no consensus standard as of today. No widely adopted metric specifically for preservation of important details or semantic fidelity in production. The closest things are embedding based semantic metrics and LLM as judge approaches, and both are still research stage.



Corn

Which is the honest state of things, and it also tells you how hard the problem is. Everyone knows the standard metric is wrong, and nobody has agreed on the replacement.



Herman

Code switching is the case that makes the whole problem visible. ElevenLabs Scribe came out with the lowest WER in one commercial comparison at thirteen point two overall, and a BERTScore of zero point nine three six. The paper's conclusion about the pattern is that WER inflates the magnitude of quality gaps by about three times. Another one, on English and Yoruba code switched speech, puts it flatly. Aggregate WER hides code switching behavior.



Corn

Because a transcript that switches language mid-sentence is being scored against a reference that has its own opinion about how to write that switch.



Herman

And dysarthric speech makes the same point from the other end. Across commercial systems tested on severe dysarthria, every single one came in above fifty-one percent WER. On mild dysarthria the same systems land at one to two percent. That's not a curve, that's a cliff. And for someone with severe dysarthria, the transcript that matters is the one a person can still read.



Corn

Now, benchmarks. How are these test sets actually built.



Herman

The standard ones. LibriSpeech is audiobook read speech, and it's everybody's favorite because the audio is clean and the speakers are cooperative. FLEURS is Wikipedia sentences read by native speakers, a hundred and two languages. Common Voice is crowd sourced volunteer recordings, so quality varies by contributor. Multilingual LibriSpeech. VoxPopuli is European Parliament. All of them are read or semi-prepared speech.



Corn

And then people test a dictation app on actual human beings talking over each other.



Herman

Maheshwari and co-authors ran exactly that comparison. Canary scores WER zero point one nine on LibriSpeech. On TalkBank conversational phone calls it scores zero point five four. Whisper goes from zero point one one on LibriSpeech to zero point five four on the conversational segments. And on the variant where speakers switch mid-recording, Whisper comes in at one point seven three.



Corn

From eleven percent to a hundred and seventy-three percent. Same model.



Herman

And they found a positive correlation between conversational markers, the coughs and the laughter and the interruptions, and WER. Which means the benchmark isn't measuring the model's transcription ability so much as its tolerance for real speech. LibriSpeech doesn't have people interrupting each other.



Corn

There's a contamination problem on top.



Herman

The Swiss German study again. Published state of the art results sit at seventeen point one to seventeen point five percent WER. They then took a vanilla Whisper model, gave it zero Swiss German data, and had it train on the ASGDTS test set itself. It hits thirteen point eight eight percent. It beats every published system.



Corn

It beat them by memorizing the test.



Herman

Phi four multimodal got to three point nine percent WER purely by memorization. Their conclusion is that the benchmark primarily measures convention matching rather than dialectal comprehension.



Corn

The leaderboard for that dialect was measuring who had seen the answers.



Herman

Which is the thing that should make you suspicious of any SOTA WER claim. Not that the work is dishonest, but that the number is achievable by routes that have nothing to do with understanding speech.



Corn

There are data quality issues in the multilingual sets too.



Herman

ACL work last year on multilingual datasets found extremely short duration utterances, inadequate quality control, and sociolinguistic blind spots. You can have thousands of hours of speech and very little of it be representative.



Corn

If I'm making a deployment decision, what do I actually do. I can't run a human evaluation on every hour of audio.



Herman

Use WER and CER as a cheap first filter. They're good at that. They'll separate a system that's broken from a system that's in the ballpark. But for anything you're going to ship, run a small human evaluation on audio that looks like your actual audio. Not LibriSpeech. Your calls, your clinic, your meeting recordings.



Corn

Check for hallucinations specifically on the near perfect transcripts. The ones where you'd otherwise trust the score.



Herman

That's the counterintuitive one. The clean transcript is where you have to look, because the metric will never flag it. And evaluate readability separately from accuracy. Those come apart. The Hugging Face course shows the same audio through two systems. Wav2Vec2 gives you all caps, no punctuation, and 'CHRISTMAUS' and 'SIMALYIS' spelled wrong. Whisper gives you readable orthographic output. Same content, wildly different usefulness to a person.



Corn

Liao and co-authors had the line on that. A perfectly accurate transcript can still be challenging to read because of grammar errors, disfluency, and the other errata that are just part of spoken communication.



Herman

That's before you get to speaker attribution. Multi speaker audio needs speaker attributed WER, and diarization error rate for who spoke when. The diarization metric is itself sensitive to pause annotation ambiguity, which can inflate the error rate or hide a genuine model error. So even the secondary metric has a measurement problem inside it.



Hilbert

It's the middle of the word that gets you, not the whole word.



Corn

How's that.



Hilbert

My niece. She's a transcriptionist for a medical billing company. Has been for six years. She edits ASR drafts all day before they go anywhere near a bill. She says the drafts are mostly right and completely useless. The model gets every drug name slightly wrong, and it gets it wrong in a way that still looks like a drug name.



Corn

Give me an example.



Hilbert

Metformin and metronidazole. Hypertension and hypotension. One substitution each. WER scores it the same as 'sit' for 'sat'. She keeps a running list. She calls the entries on it the expensive ones.



Corn

Because hypertension and hypotension are opposites.



Hilbert

One of them you take a pill for. The other one you go to the emergency room for. She's had two cases where an editor caught it on the way past and one where the editor was new.



Corn

She brought this to management.



Hilbert

She brought the list. They'd bought the system off a spec sheet that said ninety-five percent accurate. Nobody at the company knew what WER was. They read ninety-five percent and assumed ninety-five percent of the transcript was correct. They'd budgeted for one part time editor. They hired two more.



Corn

The spec sheet was true and it misled them anyway.



Hilbert

True and useless, same as the drafts. Then there's the banana bread.



Herman

The what.



Hilbert

A chart came through with an entire section of a patient's history replaced by a recipe for banana bread. Quantities, oven temperature, the instruction to let it cool. Perfectly coherent. Sat in the middle of the record like it belonged there. The model hallucinated it out of some burst of noise.



Corn

Somewhere in the model, a loaf.



Hilbert

She flagged it and pulled it. But she also made it, because the recipe was in front of her and she was curious. Brought it to Thanksgiving. It's good. The family calls it the ASR banana bread now.



Corn

The hallucination that defeated the metric produced a usable recipe and an unusable medical record.



Hilbert

That's about the size of it. There's a family in Queens that asks her to make it every year.



Corn

The number on the box said ninety-five percent.



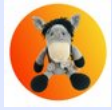
Herman

The number was correct. It just wasn't measuring the thing they thought it was measuring.



Corn

Here's where I've landed. If the metric that agrees with human judgment the least is the one everybody reports, the question isn't whether to replace it. It's what would have to be true for WER to lose the default slot. It's free, it's reproducible, and it's comparable across a decade of systems. Those aren't small things.



Herman

Whatever replaces it has to be as cheap and as portable, and none of the candidates are. SemDist, EmbER, POSER, the sentence level semantic error rate, the NLI based intelligibility work, SHALLOW, HALAS. They're all research stage. Nobody has a number you can put on a spec sheet.



Corn

The gap between the benchmark and the room is widening. A model that learns the test set beats every published system on it. A transcription tool is accurate and useless in the same sentence. Those two facts are the same fact.



Herman

Herman Poppleberry, and our producer Hilbert Flumingtop, who has a recipe now.



Hilbert

I've had the recipe for two years.



Corn

If you enjoyed this one, try episode three, Safetensors or something else; and episode eight, Building Your Own Whisper. This has been My Weird Prompts. If you want to send us something to chew on, the bot's on Telegram at t dot me slash MWP listener bot. We'll be back soon.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **HuggingFace Audio Transformers Course, ch. 5 Evaluation metrics for ASR [primary]**
   https://github.com/huggingface/audio-transformers-course/blob/main/chapters/en/chapter5/evaluation.mdx

2. **Wikipedia, Word error rate**
   https://en.wikipedia.org/wiki/Word_error_rate

3. **TorchMetrics, WordErrorRate docs**
   https://torchmetrics.readthedocs.io/en/v1.1.2/text/word_error_rate.html

4. **Manohar et al., What is lost in Normalization? Exploring Pitfalls in Multilingual ASR Model Evaluations, arXiv:2409.02449v3 (2024)**
   https://arxiv.org/html/2409.02449v3

5. **Thennal D K et al., Advocating Character Error Rate for Multilingual ASR Evaluation, NAACL 2025 Findings, arXiv:2410.07400**
   https://aclanthology.org/2025.findings-naacl.277/

6. **Sharma et al., Rethinking Human-Aligned Evaluation: An Analysis of Semantic Metrics Beyond WER, arXiv:2609.21663v2 (2026)**
https://arxiv.org/abs/2609.21663v2

7. **McNamara et al., Style-agnostic evaluation of ASR using multiple reference transcripts, arXiv:2412.07937v1 (2024)**
https://arxiv.org/abs/2412.07937v1

8. **Akeret, Subtitle-Aligned Fine-Tuning of Whisper for Swiss German ASR, arXiv:2606.07608v1 (2026)**
https://arxiv.org/abs/2606.07608v1

9. **Maheshwari et al., ASR Benchmarking: Need for a More Representative Conversational Dataset, arXiv:2409.12042v1 (2024)**
https://arxiv.org/html/2409.12042v1

10. **Barański et al., HALAS: A Human-Annotated Dataset of Hallucinations of Modern ASR Systems, arXiv:2606.23048v1 (2026)**
https://arxiv.org/abs/2606.23048v1

11. **Koudounas et al., Hallucination Benchmark for Speech Foundation Models (SHALLOW), arXiv:2510.16567v1 (2025)**
https://arxiv.org/abs/2510.16567v1

12. **Bañeras-Roux et al., Evaluation of Automatic Speech Recognition Using Generative Large Language Models, arXiv:2604.21928v3 (2026)**
https://arxiv.org/abs/2604.21928v3

13. **Phukon et al., Aligning ASR Evaluation with Human and LLM Judgments, arXiv:2506.16528v2 (Interspeech 2025)**
https://arxiv.org/abs/2506.16528v2

14. **Abdoli et al., Benchmarking Commercial ASR Systems on Code-Switching Speech, arXiv:2605.19069v3 (2026)**
https://arxiv.org/abs/2605.19069v3

15. **Okocha & Grant, Beyond Word Error Rate: A Switch Aware Evaluation... English Yoruba Code-Switched Speech, arXiv:2609.11786v1 (2026)**
https://arxiv.org/abs/2609.11786v1

16. **Altinok, Mind the Gap: Entity-Preserved Context-Aware ASR Structured Transcriptions, arXiv:2506.22858v1 (2025)**
https://arxiv.org/abs/2506.22858v1

17. **Alsayegh & Masood, Benchmarking Commercial Speech Recognition... on Dysarthric Speech, arXiv:2512.17474v2 (2026)**
https://arxiv.org/abs/2512.17474v2

18. **Waheed et al., On the Robust Approximation of ASR Metrics, arXiv:2502.12408v2 (ACL 2025)**
https://arxiv.org/abs/2502.12408v2

19. **Horiguchi et al., Diarization Error Decomposition Under Pause Annotation Ambiguity, arXiv:2609.11007v2 (2026)**
https://arxiv.org/abs/2609.11007v2

20. **Datatoorder, Character Error Rate (CER)**
https://datatoorder.com/technical/character-error-rate/

21. **Abbasihafshejani & Jadliwala, Reducing Hallucinated Transcripts in Whisper via Hallucination Space Projection, arXiv:2609.04561v1 (2026)**
https://arxiv.org/abs/2609.04561v1