LASER: An LLM-based ASR Scoring and Evaluation Rubric

Amruta Parulekar **Preethi Jyothi**
Indian Institute of Technology Bombay, Mumbai, India
amrutaparulekar.iitb@gmail.com, pjyothi@cse.iitb.ac.in

Abstract

Standard ASR evaluation metrics like Word Error Rate (WER) tend to unfairly penalize morphological and syntactic nuances that do not significantly alter sentence semantics. We introduce an LLM-based scoring rubric LASER that leverages state-of-the-art LLMs' in-context learning abilities to learn from prompts with detailed examples. Hindi LASER scores using Gemini 2.5 Pro achieved a very high correlation score of 94% with human annotations. Hindi examples in the prompt were also effective in analyzing errors in other Indian languages such as Marathi, Kannada and Malayalam. We also demonstrate how a smaller LLM like Llama 3 can be finetuned on word-pair examples derived from reference and ASR predictions to predict penalty types with close to 89% accuracy.

1 Introduction and Related Work

Automatic Speech Recognition (ASR) is used in a variety of applications ranging from voice assistants (Schwarz et al., 2023) to accessibility aids (Green et al., 2021). This makes it increasingly important to design accurate evaluation metrics for ASR systems. The most widely used ASR evaluation metric is word error rate (WER) (or character error rate, i.e., CER, for languages that do not have well-defined word boundaries). WER/CER are edit distance-based metrics that compute the minimum number of substitutions, insertions and deletions needed to transform an ASR prediction to its corresponding reference transcription. Lexically-sensitive metrics that are based on exact matches like WER penalize a prediction even if the ASR error is very minor in nature. This limitation of WER gets further amplified for Indian languages.

Several characteristics of Indian languages render WER a sub-optimal ASR evaluation metric: 1. Many Indian languages are morphologically rich with words having many inflectional variants (Vikram, 2013), words containing gender/tense/number markers (Pitale and Sarma, 2013), words being agglutinative in nature (Krishnamurti, 2003), etc. ASR predictions might contain minor errors in terms of these morphological inflections which get treated as major errors by WER. 2. Compound words are common across many Indian languages (Kulkarni et al., 2012). There are multiple accepted forms of writing the same word (e.g., *paas wala* vs. *paaswala*). WER treats one form as an error if the reference contains the other. 3. Many Indian languages contain English loan words that do not have standardized native script spellings. (E.g., ice cream in Devanagari could be written as *ayskrim* or *aaiskreem*). Although such variants should be treated the same, WER penalizes any variant differing from the reference. There are other error types common to English and Indian languages like colloquialisms (dunno vs. don't know), abbreviations (brb vs. be right back), numerical phrases (10 vs. ten), etc. that should also ideally incur no penalty during evaluation.

Semantic metrics like BERTScore (Zhang et al., 2020) or SemDist (Kim et al., 2021) are based on embeddings and do not always fare well on alternate spellings. Phoneme Error Rate (PER) (Yolchuyeva et al., 2019) and CER accommodate alternate phoneme/character-based spellings but treats all errors equally, regardless of semantic impact. Thus, there is a need for a nuanced evaluation metric that heavily penalizes semantically significant errors, lightly penalizes minor ones, and ignores acceptable variations. Large language models (LLMs), with their strong in-context learning abilities, can be leveraged for this purpose.

In this work, we propose a novel LLM-based scoring and evaluation rubric for ASR (LASER). LASER avoids penalizing colloquial spelling variations, compound words, alternate transliteration

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 24774–24783
November 4-9, 2025

Error type	Example variations	Penalty
Numerical Phrases	"1300" vs "Terah sau" or "Ek hajar teen sau"	No penalty
Abbreviations	"ATM" vs "Ay Ti Em" vs "Ay tee yum"	No penalty
Compound Words	"bhajan sangraha" vs "bhajansangraha"	No penalty
Transliterations (Native spellings)	"ayskreem" vs "aaiskrim" or "skul" vs "skool"	No penalty
Actual transliterations	"ice cream" vs "ayskrim" or "aaiskrim"	No penalty
Acceptable alternate spellings	"sundar with a bindu" vs "sundar with a half na"	No penalty
Proper nouns	"Priya" vs "Pria" vs "Preeya" vs "Preya"	No penalty
Slang and Colloquial terms	"Yaha" vs "Ye" or "vaha" vs "vo" or "par" vs "pe"	No penalty
Small (single character) spelling errors	"ladki" vs "ladkee" or "bahut" vs "bahoot"	Minor penalty
Small grammatical errors (gender/tense/number)	"hain" vs "hai" or "uska" vs "uski" vs "usko"	Minor penalty
Spelling errors that alter meaning	"kumar" vs "kamar" or "saman" vs "samanya"	Major penalty
Incorrect word substitutions	"sundar" vs "bhadda" or "mota" vs "chhota"	Major penalty
Significant omissions or additions	"–" vs "sundar" or "mota" vs "–"	Major penalty
Reordering of words that changes meaning	"bahut accha khana" vs "bahut khana accha"	Major penalty

Table 1: Types of ASR errors and their penalties.

spellings, and variant representations of numbers and abbreviations. It applies minor penalties to spelling or grammatical errors that preserve sentence meaning, and major penalties to word insertions, omissions, and meaning-altering errors. This was achieved via a carefully curated prompt to state-of-the-art LLMs and the LLM scores were compared with scores from humans who were given the same instructions. LASER correlates very well with human scores, unlike WER. Cross-lingual tests assessed whether multilingual LLMs could transfer this knowledge from high- to low-resource languages; interestingly, a prompt with Hindi examples transfers well to other Indian languages like Marathi, Kannada and Malayalam. Finally, we tested whether an open-source LLM could be trained on a word-pair dataset curated using LLM outputs to our prompt, to make our metric and penalization strategy publicly available.[1] LASER is an open-source, LLM-based fine-grained scoring metric for ASR, which attains high agreement with human evaluations.

In recent work, Tomanek et al., 2024 designed LATTEScore, an LLM-based metric that assesses meaning preservation in ASR transcripts of impaired speech through classification of sentences based on whether their meaning is preserved. Phukon et al., 2025 is concurrent work aligned with our primary objective; it combines natural language inference scores with semantic similarity and phonetic similarity to evaluate logical similarity between the prediction and ground truth. However, it focuses on correctability of dysarthric English speech while we focus on a metric that accounts for linguistic nuances of different languages.

[1]Our scripts and checkpoints to use LASER for Hindi available at https://github.com/Amparulekar/LASER-metric

2 Methodology

Metric development. To develop LASER, we first analyzed error types penalized by standard ASR metrics and assigned revised penalties: lower for minor grammatical errors and higher for semantic errors. Semantically equivalent variations, like compound words and transliterations, should incur no penalty. Table 1 lists the no-penalty, minor-penalty, and major-penalty error types that we identified. Major and minor errors were assigned penalties of 1 and 0.5 points, respectively. Non-penalizable errors incurred no penalty. A sentence-level score is then defined as $1 - \frac{\text{Total penalty}}{\text{Number of reference words}}$.

Prompt development. Our main LLM prompt (Appendix A) is in English (Latin script) with examples in Hindi (Devanagari script). Our prompt has three main components:

(a) Detailed instructions: The LLM is instructed to tokenize sentences by words, align ground-truth labels with predictions forming word pairs, and identify mismatches. The LLM is further instructed to classify the mismatches by error type via detailed examples and assign major, minor or no penalty, and finally, compute the sentence-level LASER score by adding up the penalties.

(b) Detailed examples: We provided an example for every error type (shown in Table 1).

(c) Promote step-by-step reasoning: We promoted chain-of-thought reasoning by asking the LLM to return its response in the format: *(Word count of original sentence; list of non-penalizable errors; list of major penalizable errors; list of minor penalizable errors; total penalty; score)*. This format ensured that the LLM applied the metric consistently.

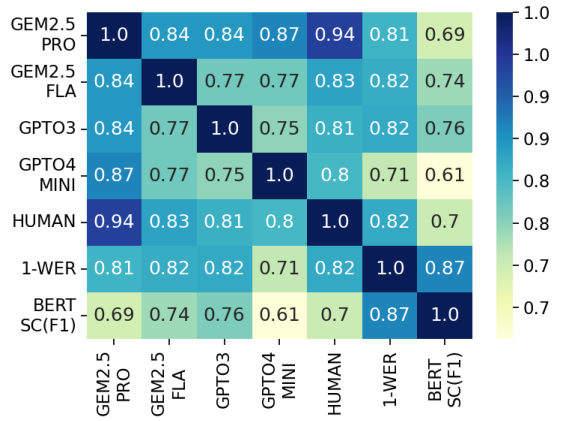
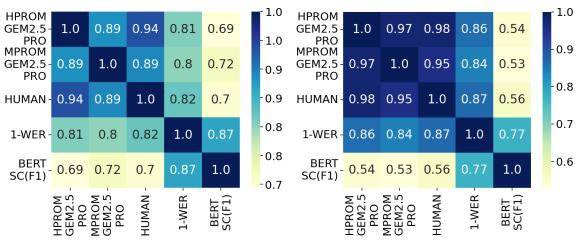


Figure 1: Correlation heatmap for different LLM scores using the Hindi prompt, Human scores, WER and BERTScore(F1) on Hindi data.



(a) Hindi data (b) Marathi data

Figure 2: Correlation heatmap for Human scores, WER, BERTScore(F1) and Gemini 2.5 Pro scores using the Hindi *(HPROM)* and the Marathi *(MPROM)* prompts for both the Hindi and the Marathi data.

Dataset creation for LLM Finetuning. Since LLM API calls are expensive, we investigated whether we could finetune a smaller LLM (e.g., Llama 3.1) with aligned word pairs (derived automatically by aligning the ASR output and the reference) to predict whether the word-pair incurs a major, minor or no penalty. Word pairs for training are obtained via human-annotated transcripts. Section 4.4 provides more details of this experiment.

3 Experimental Setup

Dataset. We used a subset of the IndicVoices test set (Javed et al., 2024), a multilingual, multi-speaker collection of natural, spontaneous speech. We used the multilingual SeamlessM4T (Communication et al., 2023) model to generate ASR predictions. Sentence pairs with no transcription mismatch (i.e., 0 WER) were removed. We focused on two Indo-Aryan (Hindi, Marathi) and two Dravidian (Malayalam, Kannada) languages. Our final datasets had 172 Hindi, 154 Marathi, 229 Malayalam, and 216 Kannada sentence pairs.

LLMs. For prompt-tuning, we chose LLMs known for their strong reasoning capabilities. From the Gemini (Team et al., 2024) and GPT (OpenAI et al., 2024) families, Gemini 2.5 Pro and GPTo3 are advanced reasoning models, while Gemini 2.5 Flash and GPTo4mini offer speed and cost-efficiency.[2] For finetuning of the word-pair classification task, we chose the open-source Llama 3 8B. (Grattafiori et al., 2024).

Score evaluation. To evaluate LLM outputs, we assigned the same task to humans using identical instructions, examples, and a worked-out sample. The humans were paid Rs. 24 per sentence to list sentence-wise major, minor and non-penalizable errors and their counts; more details are in Appendix B. These penalty counts were used to calculate our LASER scores. We computed the Pearson's correlation coefficients between human, LASER, and standard metric scores. A higher human-LASER score correlation compared to human-WER correlation would indicate that the LASER scores are more accurate.

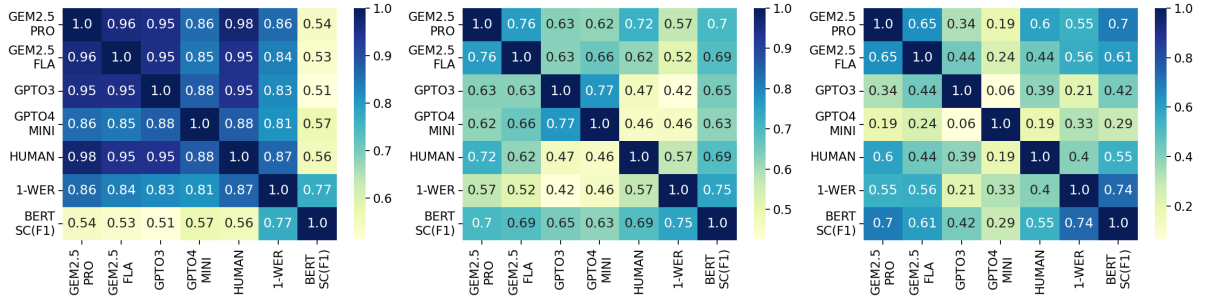
4 Experiments and Results

4.1 Correlation analysis

Figure 1 depicts the correlation heatmap of LLM-based LASER scores, human scores and standard metrics (WER, BERTScore) for 172 Hindi sentences. Gemini 2.5 Pro outperformed every other LLM, exhibiting the highest correlation with human scores and significantly surpassing WER correlation scores. Additionally, Gemini 2.5 Pro consistently recalled the initial scoring instructions, showed clear reasoning, and formatted results correctly after each batch. Notably, it was also able to infer appropriate penalties for error types not included in the prompt, viz. sandhi (phonetic transformation at word boundaries during word fusion) (Dave et al., 2020) and synonyms.

IndicVoices is primarily noisy and contains a majority of spontaneous/conversational speech. It is also rich in dialectal diversity with speech covering 145 Indian districts & 22 languages. Our experiments demonstrate that LASER performs well on noisy & dialectal speech.

[2]Deepseek R1 failed to produce scores using the prompt.



(a) Marathi (b) Malayalam (c) Kannada

Figure 3: Correlation heatmaps for different LLM scores using the Hindi prompt, Human scores, WER and BERTScore(F1) for 154 Marathi, 229 Malayalam and 216 Kannada sentences.

Lang	Categ	Predicted	Original	WER	OUR	HUMAN	Types of errors
Hindi	High WER high score	मुझे मेरा फिन केयर स्माल फाइनेंस बैंक खाता शेष और लेनदेन इतिहास दिखाएँ *mujhe mera fin keyar smal fainens bank khata shesh aur lenden itihas dikhaein*	मुझे मेरा फ़िन्कैर इस्मौल फाइनैन्स बैंक खाता शेष और लेनदेन इतिहास दिखाएं *mujhe mera finkair ismaul fainains bank khata shesh aur lenden itihas dikhaen*	0.4167	1	1	Technical terms, Compound words, Transliterations, Alternate spellings
		न्यू यॉर्क स्पेन रूस लंडन और दुबई *Nyoo York spen roos (russia in hindi) londdon - -*	न्यूयॉर्क स्पेन रशिया लंदन और दुबई *Nyooyork spen rushia london aur dubai*	0.6667	1	0.9167	Proper nouns, Transliterations, Compound words, Translations
Marathi	High WER high score	आणि त्याची सर्व्हिसही योग्य प्रकारे ती देत नाही *ani tyachi sarvhisahi yogya prakare ti det nahi*	आणि त्याची सर्विसही योग्यप्रकारे ते देत नाही *ani tyachi sarvisahi yogyaprakare te det nahi*	0.5714	0.9286	0.9286	Transliterations, Compound words, minor grammar errors
		हो हो हो हो हो हो हो *ho ho ho ho ho ho ho*	हां हां हां हां हां हां हां *han han han han han han han*	1	0.875	0.875	Non-verbal vocalizations, Acceptable alternate spellings
Hindi	High WER low score	हाँ अजन्नी मटर और मटर गैरेबी के साथ *haan ajanni matar aur matar gairebi ke sath*	हाँ मटन और मटन ग्रेवी साथ *haan matan aur matan grevi saath*	0.8333	0.5	0.3333	Incorrect word substitutions, Word additions, Major semantic changes
Marathi	High WER low score	ऑफर आलं की तुम्हाला लगेच चळवळी होते मग *ofar ala ki tumhala lagech chalvali hote mag*	ऑफर आले की तुम्हाला लगेच कळवते मग *ofar ale ki tumhala lagech kalavte mag*	0.4286	0.6429	0.7143	Incorrect word substitution, Word addition, Major semantic changes

Figure 4: Qualitative analysis of high WER samples having high and low LASER scores. Red text indicates mismatch between the original and predicted transcriptions.

4.2 Cross-lingual transfer

We developed a new prompt having English instructions with Marathi examples similar to the Hindi prompt examples. To compare the efficacy of cross-lingual transfer between higher- (Hindi) and lower resource (Marathi) languages of the same language family (Indo-Aryan), we used the Marathi and the Hindi prompts on both the Hindi and the Marathi sentences. Only Gemini 2.5 Pro was used, as it was the best-performing LLM for both languages. Figures 2a and 2b show the correlations using both prompts on the Hindi and the Marathi sentences respectively. Although both prompts gave high human score correlations, the Hindi prompt performed better for both languages, likely due to the LLM's familiarity with Hindi.

To evaluate cross-lingual inference, i.e., whether the LLM could infer mismatch types in a new language given examples of a different language, the prompt with Hindi examples was used to obtain scores for Marathi, Malayalam and Kannada. Malayalam and Kannada are lower-resource Dravidian languages that are syntactically and morphologically very different from Hindi. The scoring method and the LLMs were the same as those for Hindi. Figures 3a, 3b and 3c depict the correlation heatmaps of Hindi-prompted LLM scores, standard metrics and human scores for the three languages. The Hindi prompt transferred effectively and we observe trends similar to Hindi. Gemini 2.5 Pro was the best performing LLM, yielding the highest correlation with human scores. This indicates that LLMs like Gemini 2.5 Pro are able to adapt grammar rules from one language to another (even from a different language family). The correlation scores for Marathi were higher than those for Hindi; this could be due to the shorter average sentence length (20.86 words for Marathi vs. 27.34 words for Hindi) which may have reduced processing complexity. The correlation scores for Malayalam and Kannada were lower than those for Hindi and Marathi, potentially due to nuances of Dravidian languages that the Hindi prompt was unable to address. Notably, even on Dravidian languages, Gemini 2.5 Pro was able to consider target-language nuances beyond those explicitly included in the Hindi prompt. This experiment demonstrates that our carefully designed prompt can be scaled to multiple languages. We also ran an experiment on English, details of which are in Appendix C.

4.3 Qualitative Analysis

We qualitatively analyzed high WER samples (greater than 0.35) that had high and low LASER scores. On instances with high WERs, we checked whether our metric does indeed correct the unfair penalization of semantically identical but syntactically different mismatches between references and ASR predictions. Figure 4 shows how high WER and high LASER score samples contained a high percentage of non-penalizable errors of different types (while retaining the word-pair meanings); in contrast, low LASER score samples had significant semantic word-pair mismatches. Moreover, human scores are consistent with the LASER scores. This validates the necessity and utility of our metric.

4.4 Finetuning for word-pair classification

To develop a more efficient way to use our metric, we performed low-rank adaptation (LoRA) finetuning of the Llama3-8B model on a word-pair classification objective for the classes - "No mismatch", "Non-penalizable error", "Major penalty" and "Minor penalty". The LoRA model contains 3.4M trainable parameters. We used 950 word-pairs to finetune this model.

Evaluation was done in two ways: 1) test-train split of the train set (to evaluate classification accuracy), 2) holding out 17 out of 172 sentences as a test set prior to train set creation (to evaluate scoring efficacy). A Hindi word-pair classification dataset was curated after manual corrections to Gemini 2.5 Pro outputs and adding a random set of no-mismatch pairs.

Class	#train+val	#test	#Correct	Accuracy
0 (Identical)	310	34	32	94.12%
1 (Non-Pen)	312	35	31	88.57%
2 (Minor)	77	9	6	66.67%
3 (Major)	251	28	25	89.29%
All	950	106	94	88.69%

Table 2: Class-wise accuracies on finetuning Llama3-8.

Test-train split. Table 2 depicts the model's test set accuracies, achieving 88.69% across all classes. We observe that that the minor-penalty errors are the toughest for Llama to identify, as they are more infrequent compared to the other error categories.

Held-out sentences. The 17 held-out sentence pairs were aligned into corresponding word pairs using a custom greedy alignment script. These aligned pairs were converted into a test set to evaluate the LLM. Llama predictions were used to obtain a total penalties and corresponding scores for the 17 sentences. Figure 5 depicts the correlation heatmap of Llama3 scores, Gemini-2.5-Pro scores, human scores and standard metrics for the 17 held out sentences. Llama performs even better than Gemini-2.5-Pro and is more aligned with human scores, potentially due to the manual corrections of the Gemini-2.5-Pro outputs prior to training.

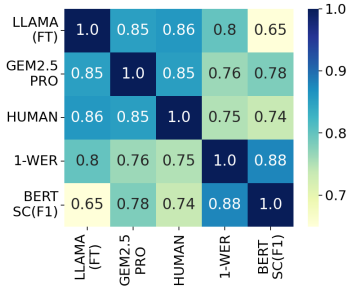


Figure 5: Correlation heatmap for finetuned Llama3, Gemini 2.5 Pro (Hindi prompt), Human, WER and BERTScore(F1) for 17 held-out Hindi sentence pairs.

5 Conclusion

In this work, we develop LASER, a fine-grained LLM-driven ASR metric that considers semantic, linguistic and morphological nuances and does not unfairly penalize predicted transcriptions. We use a carefully curated prompt with detailed descriptions of error types in Hindi. We tested the prompt on multiple LLMs and compared the results with human evaluations. We observe that LLMs like Gemini-2.5-Pro are very well-correlated with human annotations unlike standard measures like WER. We are also able to use the prompt with Hindi examples to effectively transfer knowledge to transcriptions in other languages from the same as well as different language families (Marathi, Kannada, Malayalam). Finally, we show the feasibility of a more efficient evaluation setup by finetuning Llama-3 to learn our penalty rules using a small amount of hand-annotated data.

Limitations

LLMs tend to process ambiguities differently on different runs. For instance, a slang spelling might be considered a spelling error and penalized incorrectly. It was observed that these differences were higher in case of lower-resource languages. Although this occurs in a small number of cases and the variation is small, there is a need to develop a standardized technique that will ensure the same score on all runs. Finetuning Llama with Gemini predictions addresses the LLM inference inconsistency issue, as the weights can be fixed at inference time to obtain consistent outputs.

Our prompt-based technique has higher latency compared to other metrics. We improved the efficiency of LASER through Low-rank adaptation (LoRA) finetuning of Llama, but reducing latency further can be a direction of future research.

References

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, and 49 others. 2023. Seamlessm4t: Massively multilingual & multimodal machine translation. *Preprint*, arXiv:2308.11596.

Sushant Dave, Arun Kumar Singh, Prathosh A. P., and Brejesh Lall. 2020. Neural compound-word (sandhi) generation and splitting in sanskrit language. *Preprint*, arXiv:2010.12940.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jordan R. Green, Bob MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn Ladewig, Jimmy Tobin, Michael Brenner, Philip Q Nelson, and Katrin Tomanek. 2021. Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases.

Tahir Javed, Janki Atul Nawale, Eldho Ittan George, Sakshi Joshi, Kaushal Santosh Bhogale, Deovrat Mehendale, Ishvinder Virender Sethi, Aparna Ananthanarayanan, Hafsah Faquih, Pratiti Palit, Sneha Ravishankar, Saranya Sukumaran, Tripura Panchagnula, Sunjay Murali, Kunal Sharad Gandhi, Ambujavalli R, Manickam K M, C Venkata Vaijayanthi, Krishnan Srinivasa Raghavan Karunganni, and 2 others. 2024. Indicvoices: Towards building an inclusive multilingual speech dataset for indian languages. *Preprint*, arXiv:2403.01926.

Suyoun Kim, Abhinav Arora, Duc Le, Ching-Feng Yeh, Christian Fuegen, Ozlem Kalinli, and Michael L. Seltzer. 2021. Semantic distance: A new metric for asr performance analysis towards spoken language understanding. *Preprint*, arXiv:2104.02138.

Bhadriraju Krishnamurti. 2003. *The Dravidian Languages*. Cambridge Language Surveys. Cambridge University Press.

Amba P. Kulkarni, Soma Paul, Malhar A. Kulkarni, Anil Kumar, and Nitesh Surtani. 2012. Semantic processing of compounds in indian languages. In *International Conference on Computational Linguistics*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

B. Phukon, X. Zheng, and M. Hasegawa-Johnson. 2025. (2025) aligning ASR evaluation with human and LLM judgments: Intelligibility metrics using phonetic, semantic, and NLI approaches. *Proc. Interspeech*, 2025:5708–5712.

Shalmalee Pitale and Vaijayanthi M. Sarma. 2013. *Marking plurals: the acquisition of nominal number inflection in Marathi*, page 99–109. Cambridge University Press.

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

Andreas Schwarz, Di He, Maarten Van Segbroeck, Mohammed Hethnawi, and Ariya Rastrow. 2023. Personalized predictive asr for latency reduction in voice assistants. *Preprint*, arXiv:2305.13794.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 32 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Katrin Tomanek, Jimmy Tobin, Subhashini Venugopalan, Richard Cave, Katie Seaver, Jordan R. Green, and Rus Heywood. 2024. Large language models as a proxy for human evaluation in assessing the comprehensibility of disordered speech transcription. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10846–10850.

Shweta Vikram. 2013. *Morphology: Indian Languages and European Languages*. International Journal of Scientific and Research Publications, Vol 3, Issue 6.

Sevinj Yolchuyeva, Géza Németh, and Bálint Gyires-Tóth. 2019. Transformer based grapheme-to-phoneme conversion. In *Interspeech 2019*, interspeech-2019, page 2095–2099. ISCA.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A LLM Prompt (Hindi)

Italicized text in the prompt below was in Devanagari script in the original instructions.

YOUR TASK:

Here's how we can address these challenges in Indian language ASR evaluation and design a scoring metric system from 0 to 1. Designing the Metric -

A) Define Non-Penalizable Errors:

1. Numbers: Accept different spellings for numbers (e.g., "1300", "thirteen hundred", "*Terah sau*").
2. Abbreviations: Accept variants in spelling abbreviations (e.g., ATM spelled as "*aytiem*" or "*ayteeyam*").
3. Compound Words: Accept variations in joining or separating compound words (e.g., "*bhajan sangraha*" vs. "*bhajansangraha*" or "*paas wala*" vs. "*paaswala*").
4. Native Spellings of Transliterated Words: Accept variants in spelling transliterations (e.g.,"*aaiskrim*" vs. "*aaiskreem*" vs. "*ayskrim*")
5. Transliterated words: Accept latin script spellings of transliterated words (e.g.,"*aiskrim*" vs. Ice cream or "*skool*" vs. School)
6. Alternate Spellings: Accept grammatically correct spelling differences (e.g., "*sundar (with a bindu)*" vs. "*sundar (with half "na")*".
7. Proper Nouns: Allow for minor spelling variations in names and places (e.g., "*priya*" vs. "*preya*" vs. "*preeya*").
8. Slang and Colloquial Terms: Account for regional variations (e.g., "*yaahaan*" vs. "*yaahaa*" or "*yaha*" vs. "*ye*").

B) Define Minor Penalizable Errors (0.5 points):

1. Small spelling error: Minor penalty for small single character spelling errors that sound similar (e.g. "*ladkee*" vs. "*ladki*")
2. Small grammatical error: Minor penalty for a small grammatical error that does not alter meaning (e.g. Gender error or singular plural error like "*uska* " vs. " *uski*" or "*hain*" vs. "*hai*")

C) Define Major Penalizable Errors (1 point):

1. Incorrect word substitutions (e.g., replacing "*sundar*" with "*bhadda*").
2. Significant omissions or additions.
3. Reordering of words that changes the meaning.
4. Spelling mistakes that change meaning (e.g. "*kumar*" vs. "*kamar*")

D) Matching Strategy:

1. Token-Based Matching: Use fuzzy matching to compare each token (word). Assign weights to tokens to prioritize penalizable errors over non-penalizable ones.
2. Phonetic Similarity: Leverage phonetic matching algorithms like Soundex or Metaphone to compare pronunciation.

E) Scoring:

1. $Score = (1 - \frac{\text{penalized errors}}{\text{total tokens}})$.
2. Weight errors differently based on severity (e.g., minor spelling variation = 0.5, word substitution = 1.0).

Once we provide the two sentences, apply the above rules and give a similarity score between 0 and 1.

EXAMPLE:

Step 1: Tokenization

Sentence 1 - predicted:
vaha, *bhajan*, *sangraha*, *komal*, *paaswala*, *aytiem*, 10, *par*, *taims*, *sundar (with bindu)*, *hain*, *skul*, *se*

Sentence 2 - original:
vo, *bhajansangraha*, *ke* , *paas*, *walaa*, A.T.M., *das*, times, *sundar(with half na)*, *hai*, *skool*, *se*

Step 2: Classify Tokens

A) Non-penalizable errors:

1. Colloquial variations:
   vaha vs. *vo*: Colloquial difference - NO PENALTY
2. Compound word handling:
   bhajan sangraha vs. *bhajansangraha*, *paaswala* vs. *paas wala*: Acceptable compound word variations - NO PENALTY
3. Abbreviation variations:
   aytiem vs. A.T.M.: Abbreviation handling - NO PENALTY
4. Numerical variation:
   10 vs. *das*: Equivalent numerical representation - NO PENALTY

5. Transliterations:
   taims vs times: Acceptable transliteration difference - NO PENALTY
6. Alternate spellings:
   isundar (with bindu) vs. *sundar (with half na)*: Regional spelling difference - NO PENALTY
7. Transliteration spelling variations:
   skul vs. *skool*: Acceptable transliteration difference - NO PENALTY

B) Penalizable errors:

1. *komal* vs. *ke*: Wrong substitution of word - Penalty weight = 1.0 (major error)
2. *par*: Addition of word - Penalty weight = 1.0 (major error)
3. *hain* vs. *hai*: Small grammatical error that changes singular to plural. - Penalty weight = 0.5 (minor error)

C) Exact matches:

1. *se*: Appears identically in both - No penalty

Step 3: Scoring

The formula is: Score $= 1 - \frac{\text{Weighted penalized errors}}{\text{Total tokens}}$
Total tokens: 12 (from Sentence 2)
Penalized errors:
komal vs. *ke* : 1.0, *par*: 1.0, *hain* vs. *hai*: 0.5
Weighted penalized errors: (1.0 + 1.0 + 0.5 = 2.5)
Score $= 1 - \frac{2.5}{12} = 1 - 0.2083 = 0.7917$
Final Similarity Score: 0.7917

STRUCTURE OF YOUR RESPONSE:

Number of tokens in original sentence; list of tokens with non-penalizable errors; list of tokens with major penalizable errors; list of tokens with minor penalizable errors; total penalty; score. If I give you predicted and original sentence pairs, can you return ONLY the output I asked for in a single json (number each sentence pair) and not details.

B Human instructions

Human annotators, who had thorough linguistic knowledge of the languages that we worked on, were commissioned to annotate our sentence pairs and obtain total penalties for each sentence. They charged us Rs. 24 per sentence pair, for the 172 Hindi, 154 Marathi, 229 Malayalam and 216 Kannada sentence pairs. Italicized text in the instructions below was in Devanagari script in the original instructions.

INSTRUCTIONS:

The annotator should look at the two sentences side by side (original and predicted) and look for any difference between the two sentences. Each difference is an error. Now of these errors, we have classified them into 3 errors, no penalty, minor penalty and major penalty. The types of errors and their classification is explained in the rules. For instance if one sentence has 1300 and the other has terah sau written, this is a no penalty error. So we need the annotators to make lists for each sentence pair, of the major, minor and no penalty errors (3 lists). Each list must be of the format '1300 vs terah sau (numbers), vaha vs vo (colloquial variation)' and so on i.e. word in first sentence vs word in second sentence and the reason why they are classified in this error type. And finally we also need the counts of the types of errors for each sentence pair.

RULES:

A) NO-PENALTY Errors:

1. Numbers: Accept different spellings for numbers (e.g., "1300", "thirteen hundred", "*Terah sau*").
2. Abbreviations: Accept variants in spelling abbreviations (e.g., ATM spelled as "*aytiem*" or "*ayteeyam*").
3. Compound Words: Accept variations in joining or separating compound words (e.g., "*bhajan sangraha*" vs. "*bhajansangraha*" or "*paas wala*" vs. "*paaswala*").
4. Native Spellings of Transliterated Words: Accept variants in spelling transliterations (e.g.,"*aaiskrim*" vs. "*aaiskreem*" vs. "*ayskrim*")
5. Transliterated words: Accept latin script spellings of transliterated words (e.g.,"*aiskrim*" vs. Ice cream or "*skool*" vs. School)
6. Alternate Spellings: Accept grammatically correct spelling differences (e.g., "*sundar (with a bindu)*" vs. "*sundar (with half "na")*".
7. Proper Nouns: Allow for minor spelling variations in names and places (e.g., "*priya*" vs. "*preya*" vs. "*preeya*").
8. Slang and Colloquial Terms: Account for regional variations (e.g., "*yaahaan*" vs. "*yaahaa*" or "*yaha*" vs. "*ye*").

B) MINOR-PENALTY Errors:

1. Small spelling error: Minor penalty for small single character spelling errors that sound similar (e.g. "*ladkee*" vs. "*ladki*")
2. Small grammatical error: Minor penalty for a small grammatical error that does not alter meaning (e.g. Gender error or singular plural error like "*uska* " vs. " *uski*" or "*hain*" vs. "*hai*")

C) MAJOR-PENALTY Errors:

1. Incorrect word substitutions (e.g., replacing "*sundar*" with "*bhadda*").
2. Significant omissions or additions.
3. Reordering of words that changes the meaning.
4. Spelling mistakes that change meaning (e.g. "*kumar*" vs. "*kamar*")

EXAMPLE:

Sentence 1 - predicted:
vaha, *bhajan*, *sangraha*, *komal*, *paaswala*, *aytiem*, 10, *par*, *taims*, *sundar (with bindu)*, *hain*, *skul*, *se*

Sentence 2 - original:
vo, *bhajansangraha*, *ke* , *paas*, *walaa*, A.T.M., *das*, times, *sundar(with half na)*, *hai*, *skool*, *se*

A) NO-PENALTY errors:

1. Colloquial variations:
   vaha vs. *vo*: Colloquial difference - NO PENALTY
2. Compound word handling:
   bhajan sangraha vs. *bhajansangraha*, *paaswala* vs. *paas wala*: Acceptable compound word variations - NO PENALTY
3. Abbreviation variations:
   aytiem vs. A.T.M.: Abbreviation handling - NO PENALTY
4. Numerical variation:
   10 vs. *das*: Equivalent numerical representation - NO PENALTY
5. Transliterations:
   taims vs times: Acceptable transliteration difference - NO PENALTY
6. Alternate spellings:
   isundar (with bindu) vs. *sundar (with half na)*: Regional spelling difference - NO PENALTY
7. Transliteration spelling variations:
   skul vs. *skool*: Acceptable transliteration difference - NO PENALTY

B) MAJOR-PENALTY errors:

1. *komal* vs. *ke* :
   Wrong substitution of word – MAJOR PENALTY
2. *par*:
   Addition of word – MAJOR PENALTY

C) MINOR-PENALTY errors:

1. *hain* vs. *hai*:
   Small grammatical error that changes singular to plural. – MINOR PENALTY

D) EXACT MATCHES:

1. *se*: Appears identically in both - Not an error

STRUCTURE OF YOUR RESPONSE:

Column 1 – list of no-penalty errors in the format (Word from sentence 1 vs word from sentence 2 (reason for no penalty), and so on)
Column 2 – list of major penalty errors in the format (Word from sentence 1 vs word from sentence 2 (reason for major penalty), and so on)
Column 3 – list of minor penalty errors in the format (Word from sentence 1 vs word from sentence 2 (reason for minor penalty), and so on)
Column 4 – number of no-penalty errors
Column 5 – number of major penalty errors
Column 6 – number of minor penalty errors

C Analysis on English

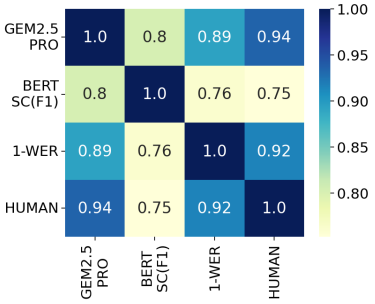


Figure 6: Correlation heatmap for LASER using Gemini 2.5 Pro (Hindi prompt), Human scores, WER and BERTScore(F1) on English data

Single example analysis

We tested LASER (Gemini 2.5 pro) for English data (using the Hindi prompt) with the utterance:

- Original: The colorful bumblebee stung unlucky Priya 3 times on the arm though.
- ASR prediction: The colourful bumble-bee strung Pria three times on the arms tho.

Categ	Predicted	Original	WER	OUR	HUMAN	Types of errors
High WER high score	it's a year-round known profit wellness camp for children and adults	it is a year-round non-profit wellness camp for children and adults	0.3636	0.9091	0.9091	Contractions
	dutzi earned 10 academy awards nominations including hoffman's 5th nomination	tootsie earned ten academy award nominations including hoffman's fifth nomination	0.4	0.85	0.95	Proper nouns, Numerical phrases, Minor grammatical
	5 winners have subsequently achieved victory on the following years' derogate	five winners have subsequently achieved victory in the following year's derby	0.3636	0.7727	0.7727	Numerical phrases, Minor grammatical errors
High WER low score	paper scales are right with much care	paper is scarce so write with much care	0.5	0.5	0.5	Incorrect word substitutions, Word omissions
	volume right down the speed	william ran down the street	0.6	0.4	0.4	Incorrect word substitutions
	missouri site	a sorry sight	1	0	0	Incorrect word substitutions, Word omissions

Figure 7: Qualitative analysis of high WER English data samples having high and low LASER scores. Red text indicates mismatch between the original and predicted transcriptions.

LLM response:

- Tokens in original: 12
- 5 No-penalty errors: "colorful vs. colourful" (Alternate spellings), "bumblebee vs. bumble-bee" (compound words), "Priya vs. Pria"(Proper nouns),"3 vs. three" (Numbers),"though vs. tho" (Slang/ Colloquial terms)
- 1 Minor penalty error: "arm vs. arms" (Small grammatical errors)
- 2 Major penalty errors: "stung vs. strung" (incorrect word substitution),"omission of unlucky"(omissions/additions)
- Total penalty: 2.5 and Score: 0.7917
- Here 1 – WER is 0.3333 (7 substitutions, 1 insertion, WER 0.6667)

We can observe that the prompt with Hindi examples transfers well to English sentences.

Quantitative analysis

Subsequently, we conducted an experiment on 80 samples of the Common-voice English test set, transcribed (Ardila et al., 2020) using the Whisper (Radford et al., 2022) model. The Hindi prompt was used with the Gemini 2.5 model to get LASER scores for English data. Figure 6 depicts the correlations on using the Hindi prompt on English data. It can be observed that LASER scores correlate the best with human evaluations, but the difference between the correlation of WER with human scores and the correlation of LASER with human scores is significantly smaller for English. This indicates that there are fewer no-penalty and minor penalty errors in the English language. This can be due to English being relatively less morphologically complex than Indic languages. Out of the 9 no-penalty error types, only 3 – slang/contractions, proper nouns and numerical phrases - are predominantly observed in English.

Qualitative analysis

We performed a qualitative analysis on English samples. Figure 7 compares samples with high WER (greater than 0.35) and high LASER scores to samples having high WER and low LASER scores. It can be observed that high WER and high LASER score samples contained a high percentage of non-penalizable errors of different types (while retaining the word-pair meanings). In contrast, high WER and low LASER score samples predominantly contained significant semantic word-pair mismatches. Thus, our metric does indeed correct the unfair penalization of semantically identical but syntactically different mismatches between references and ASR predictions. Moreover, human scores are consistent with the LASER scores, thus validating the necessity and utility of our metric.