LASER: An LLM-based ASR Scoring and Evaluation Rubric

Amruta Parulekar   Preethi Jyothi
Indian Institute of Technology Bombay, Mumbai, India

The Problem: Indian languages pose ASR evaluation challenges due to morphological richness, compound words, and non-standardized transliterations/spellings.

Our Aim: To develop a semantically-aware, fine-grained ASR evaluation metric for such languages, beyond traditional metrics, to allow for fairer ASR evaluation.

The Metric

LASER assigns 0, 0.5, and 1-point penalties to no-penalty, minor-penalty, and major-penalty error types.

Score is given by:

$$1 - \frac{\text{Total penalty}}{\text{Number of words in reference}}$$

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

The Prompt

Instructions to tokenize, align, classify errors, compute LASER (English instructions, Hindi examples) → Worked out example containing errors of each error type → Output format enforcing step-by-step reasoning: (word count; lists of no-penalty, major & minor errors; total penalty; score).

The Setup

Datasets:
- Hindi, Marathi, Malayalam, Kannada - Indicvoices data; SeamlessM4T for ASR
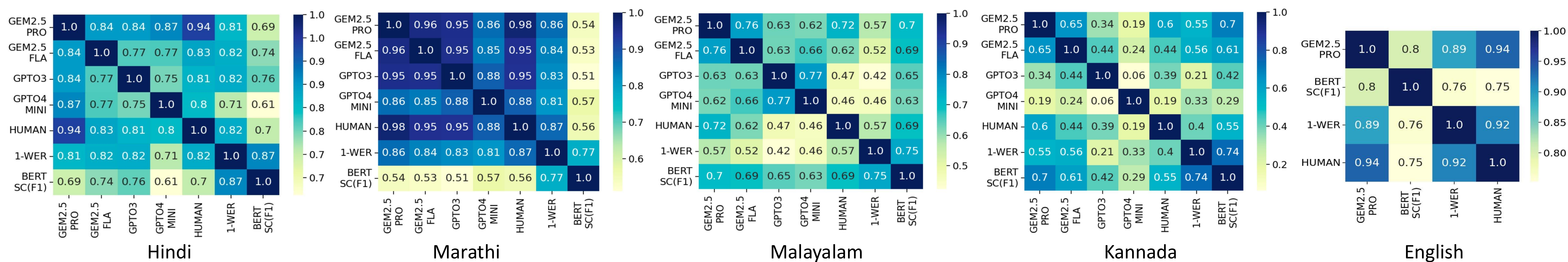- English - CommonVoice data; Whisper for ASR

LLMs:
- Gemini 2.5 Pro
- Gemini 2.5 Flash
- GPTo3
- GPTo4mini

Score Evaluation:
- Human given same instructions as the LLM prompt.
- Calculate correlation of human scores with LASER scores.

Cross-lingual Prompt Transfer
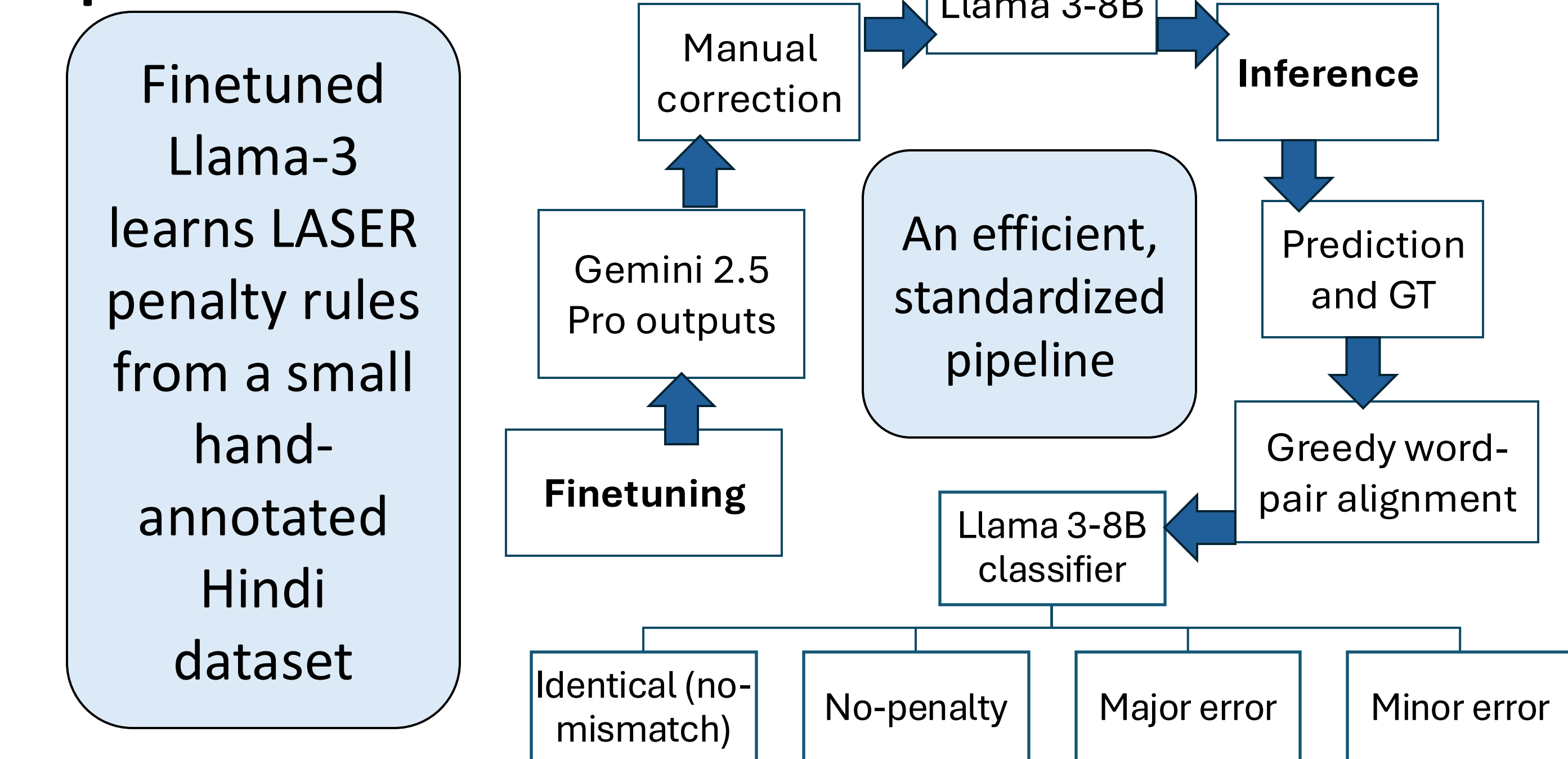

Hindi | Marathi | Malayalam | Kannada | English

- Gemini 2.5 Pro correlates more strongly with human annotations than traditional metrics do, for all languages.
- Hindi prompt transfers effectively to Marathi, Kannada, Malayalam, English (within and across language families)
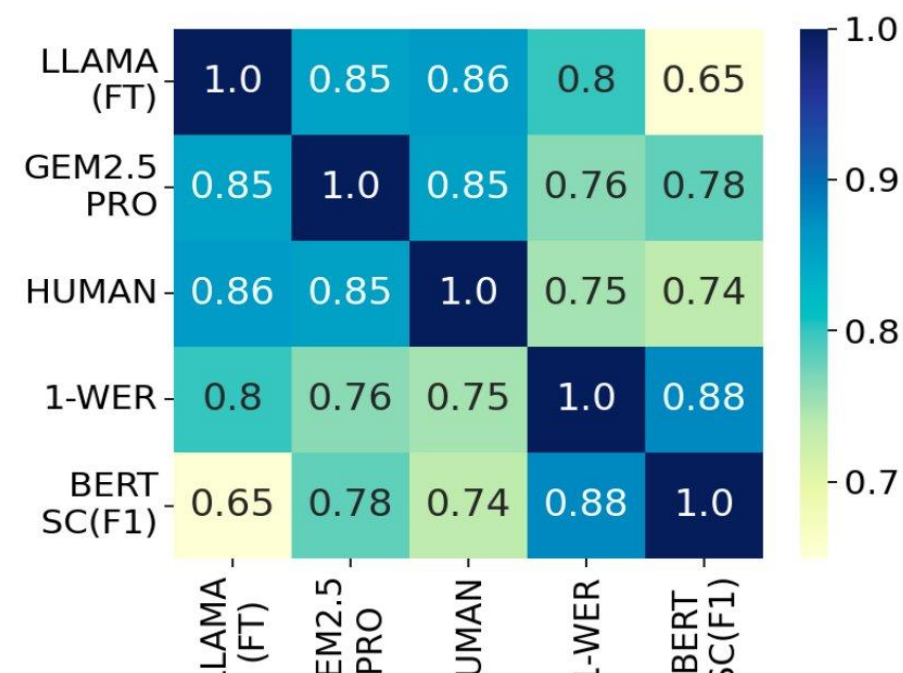
Additional observations:
- **Marathi:** A Marathi prompt and the Hindi prompt were applied to both Marathi and Hindi sentences. Both prompts correlated well with human scores, but the Hindi prompt outperformed the Marathi one—likely due to the LLM's greater familiarity with Hindi.
- **Malayalam, Kannada:** Correlation dropped due to structural differences, but language-specific nuances were captured effectively.
- **English:** The gap between WER–human and LASER–human correlations was small, indicating fewer no-penalty and minor-penalty errors—likely because English is less morphologically complex than Indic languages.

LLM Finetuning

Experiment:

Finetuned Llama-3 learns LASER penalty rules from a small hand-annotated Hindi dataset

Manual correction → Llama 3-8B → Inference → Prediction and GT → Greedy word-pair alignment → Llama 3-8B classifier → Identical (no-mismatch) | No-penalty | Major error | Minor error

Gemini 2.5 Pro outputs → Finetuning → An efficient, standardized pipeline

Results:

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%

Qualitative Analysis

High WER (>0.35) samples with high LASER scores had mostly non-penalizable errors, while low LASER scores reflected true semantic mismatches - validating LASER's effectiveness.

Categ	Lang	Predicted	Original	WER	OUR	HUMAN	Types of errors
High WER high score	Hin	फिन केयर स्माल फाइनेंस बैंक खाता शेष और लेनदेन इतिहास दिखाएँ *fin keyar smal fainens bank khata shesh aur lenden itihas dikhaein*	फ़िन्कैर इस्मौल फ़ाइनेन्स बैंक खाता शेष और लेनदेन इतिहास दिखाएं *finkair ismaul fainains bank khata shesh aur lenden itihas dikhaen*	0.417	1	1	Technical terms, Compound words, Transliterations, Alternate spellings
	Hin	न्यूयॉर्क स्पेन रूस लंडन और दुबई *Nyoo York spen roos (russia in hindi) londdon - -*	न्यॉयोर्क स्पेन रशिया लंदन और दुबई *Nyooyork spen rushia london aur dubai*	0.667	1	0.917	Proper nouns, Transliterations, Compound words, Translations
	Mar	त्याची सर्व्हिसही योग्य प्रकारे ती *tyachi sarvhisahi yogya prakare ti*	त्याची सर्विसही योग्यप्रकारे ते *tyachi sarvisahi yogyaprakare te*	0.571	0.929	0.929	Transliterations, Compound words, Minor grammar err.
		हो हो हो हो हो हो हो हो *ho ho ho ho ho ho ho*	हां हां हां हां हां हां हां हां *han han han han han han han han*	1	0.88	0.88	Non-verbal vocalizations, Alternate spellings
	Eng	it's a year-round known profit	it is a year-round non-profit	0.364	0.909	0.909	Contractions
	Eng	5 winners have subsequently achieved victory on the following years' derogate	five winners have subsequently achieved victory in the following year's derby	0.364	0.773	0.773	Numerical phrases, Minor grammatical errors
High WER low score	Hin	अजन्नी मटर और मटर गैरेबी के *ajanni matar aur matar gairebi ke*	मटन और मटन ग्रेवी *matan aur matan grevi*	0.833	0.5	0.333	Word substitutions, Additions
	Mar	आलं की तुम्हाला लगेच चळवळी होते *ala ki tumhala lagech chalvali hote*	आले की तुम्हाला लगेच कळवते *ale ki tumhala lagech kalavte*	0.429	0.643	0.714	Word substitutions, Semantic changes
	Eng	paper scales are right with	paper is scarce so write with	0.5	0.5	0.5	Substitutions, Omissions