International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

HindMed-Triage: A Hybrid Sovereign-Global AI Framework for Code-Mixed Hindi Medical Speech Triage in Rural Telemedicine

Achal Bajpai
Department of Data Science and Business Systems
SRM Institute of Science and Technology
Chennai, India
achalbajpai2004@gmail.com

Shashank Prasad
Department of Data Science and Business Systems
SRM Institute of Science and Technology
Chennai, India
shashankpd2606@gmail.com

Om Thakare
Department of Data Science and Business Systems
SRM Institute of Science and Technology
Chennai, India
om.thakare3217@gmail.com

S. Wilson Prakash
Department of Data Science and Business Systems
SRM Institute of Science and Technology
Chennai, India
wprakash.s@gmail.com

Abstract*—Rural telemedicine in India faces two compounding problems: global automatic speech recognition (ASR) systems degrade sharply on code-mixed Hindi-English patient speech, and large language models (LLMs) lack calibration for low-resource clinical triage. We propose HindMed-Triage, a seven-stage hybrid sovereign-global AI pipeline for Hindi medical speech triage, combining Sarvam Saaras V3 (sovereign ASR) with a retrieval-augmented LLM reasoning stage grounded in WHO Integrated Management of Childhood Illness (IMCI) and Indian Primary Health Centre protocols. We validate each pipeline component through systematic benchmarking. For ASR, Saaras V3 achieves WER = 0.285 on the EkaCare Hindi medical speech subset (n = 320), a 49.6% relative WER reduction over Whisper Large v3 (WER = 0.565). For triage, we construct a stratified weak-label benchmark of 500 code-mixed patient queries from MMCQSD (n_{CRITICAL} = 80) and evaluate eight LLM configurations under zero-shot and retrieval-augmented settings. GPT-OSS-120B achieves the highest zero-shot Macro F1 of 0.481 (95% CI: [0.425, 0.537]), yet bootstrap confidence intervals overlap for all eight models, indicating no statistically significant winner. A finding specific to sovereign models is that retrieval augmentation improves Sarvam-105B Macro F1 from 0.411 to 0.460, while the same augmentation reduces performance for all seven global and non-sovereign models tested, a 1-of-8 pattern that carries direct architectural implications. We further document a safety-critical failure mode: both GPT-OSS model variants send 17 to 29 of 80 true CRITICAL cases to ROUTINE under zero-shot and RAG conditions, the most dangerous possible misclassification. CRITICAL recall ranges from 0.24 to 0.51 across configurations, confirming a system-wide CONSULT-prior bias that clinical guideline injection partially corrects for sovereign models only.

***Keywords* - Code-mixed Hindi, Medical ASR, Clini- cal Triage, Sovereign AI, Retrieval-Augmented Generation, Telemedicine, Indian NLP, Weak Supervision**

I. Introduction

India's physician shortage stands at approximately 1:1,456 patients per doctor, well below the WHO-recommended 1:1,000 [1]. The shortfall is concentrated in rural areas, where 65% of the population lives but only 30% of registered physicians practice [2]. Voice-based telemedicine platforms have expanded access, yet their usefulness depends on AI capable of processing the spontaneous, code-mixed Hindi-English speech in which rural patients naturally express symptoms. A patient describing a cardiac event might say: *"bahut tez chest pain ho raha hai aur saans nahi aa rahi"*, a sentence in which Hindi grammar carries English clinical terms without belonging cleanly to either language. Existing global ASR systems such as Whisper [9] and general-purpose LLMs were not optimised for this register.

India's sovereign AI programme, led by Sarvam AI, has produced models targeting Indian languages, but no systematic external evaluation has documented where these models outperform global alternatives, where they do not, and how the two should be combined in a clinical workflow. This paper makes four contributions:

1) We propose **HindMed-Triage**, a seven-stage hybrid sovereign-global AI pipeline for code-mixed Hindi medical speech triage, covering patient speech input through structured doctor alert.
2) We benchmark ASR components (Saaras V3 vs. Whisper Large v3) on the complete EkaCare Hindi medical ASR subset (n = 320), finding a 49.6% WER advantage for the sovereign system.
3) We construct a stratified weak-label triage benchmark of 500 code-mixed patient queries and evaluate eight LLM configurations under zero-shot and retrieval-augmented settings, covering two sovereign and six global models.
4) We document a sovereign-specific RAG benefit (Sarvam-105B is the sole model improved by retrieval augmentation, 1 of 8) and a safety-critical GPT-OSS failure mode (17 to 29 CRITICAL-to-ROUTINE misclassifications per 80 true CRITICAL cases), both with direct implications for clinical deployment.

II. Related Work

A. Code-Mixed Indian Language Processing

Code-mixing in South Asian languages arises from widespread bilingualism and carries linguistic structures not

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

present in either base language [5]. Khanuja et al. introduced MuRIL, a multilingual model pre-trained on 17 Indian languages and their transliterated forms, demonstrating that India-specific pre-training yields consistent downstream gains over multilingual baselines [6]. Kakwani et al. established IndicNLPSuite, providing the corpus infrastructure from which subsequent Indian LLMs, including Sarvam-105B, derive their pre-training data [19]. Aggarwal et al. showed that code-mixed Hindi-English clinical text requires specialised tokenisation strategies, as standard subword vocabularies fragment Hindi morphology in ways that degrade semantic encoding [7]. Sitaram et al. identified medical and customer-service contexts as priority areas for code-mixed NLP research [8].

B. Multilingual and Domain-Specific Medical ASR

Radford et al. demonstrated that Whisper, trained on 680,000 hours of multilingual audio, performs well on high-resource languages but degrades substantially on low-resource and code-mixed speech [9]. This degradation is compounded in medical domains where out-of-vocabulary clinical terms increase phonetic ambiguity [10]. EkaCare published the first Indian medical ASR evaluation dataset, establishing baseline English WER values but providing no Hindi evaluation [11]. Salesky et al. showed that domain-adapted models for low-resource languages consistently outperform general-purpose models on in-domain tasks [12].

C. LLMs for Clinical Triage

Nori et al. demonstrated GPT-4's near-human USMLE performance in English, but the study did not address low-resource or non-English clinical settings [13]. Jin et al. showed that LLMs applied to clinical triage without class-frequency calibration systematically under-detect high-acuity cases due to class imbalance [15], a finding consistent with our CRITICAL-class results. Panch et al. identified label scarcity and language mismatch as the two primary failure modes of AI in resource-constrained healthcare [14]. Faujdar and Ghosh found that English-trained clinical AI systems generalise poorly to Hindi patient descriptions in Indian primary healthcare [16]. The MMCQSD dataset, used in this study, was introduced for code-mixed clinical summarisation at ECIR 2024 [17]; our triage annotation represents its first classification-oriented application.

D. Sovereign AI for Indian Languages

Sarvam AI released Sarvam-105B as India's first domestically developed 100B-parameter mixture-of-experts model trained on an Indian-language-centric corpus [18]. No independent third-party evaluation of either Sarvam-105B or Sarvam-30B existed prior to this work. The IndicBERT and IndicBART series from AI4Bharat established the viability of India-specific pre-training for classification and generation tasks [19], providing research context for evaluating the Sarvam model family.

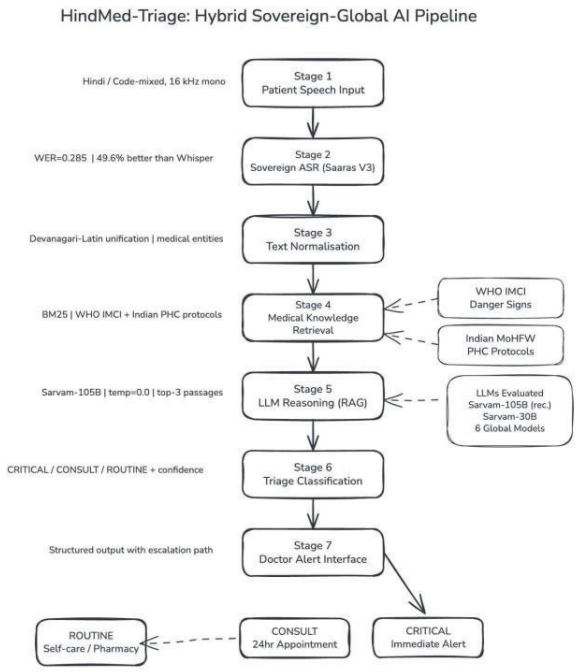

Fig. 1. The HindMed-Triage seven-stage pipeline. Stages 1 and 2 (ASR) and Stages 4 through 6 (retrieval and triage) are benchmarked in this paper. Stage 3 (normalisation) and Stage 7 (alert interface) are described as design specifications.

III. Methodology

A. The HindMed-Triage Framework

HindMed-Triage is a seven-stage modular pipeline designed for deployment in rural Indian telemedicine settings. Two design principles govern the architecture. First, sovereign components are used where empirical advantage over global alternatives is demonstrated at that stage; this paper establishes this for ASR and partially for the triage reasoning stage under retrieval augmentation. Second, zero-shot LLM CRITICAL-class failure is addressed through retrieval augmentation rather than model substitution. Fig. 1 presents the full architecture.

1) Stages 1 and 2: Speech Input and Sovereign ASR: Patient audio is sampled at 16 kHz mono. Stage 2 applies Sarvam Saaras V3 for transcription. The empirical basis for this choice is the 49.6% WER advantage over Whisper Large v3 on Hindi medical speech. In deployments where Saaras V3 API access

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

is unavailable, a local Whisper variant provides a documented fallback at WER = 0.565.

2) Stage 3: Text Normalisation: Code-mixed ASR output presents three normalisation problems. First, the same clinical concept may appear in Devanagari script, romanised transliteration, or English within the same utterance. Second, numeric medical values are attached to Hindi unit expressions. Third, Hindi negation markers invert clinical polarity. Stage 3 maps all surface forms to canonical English medical terminology using a deterministic lexicon of 30 Hindi-to-English term mappings and two regex patterns for temperature and heart rate normalisation.

3) Stage 4: Medical Knowledge Retrieval: A lightweight BM25 retrieval index [21] operates over 31 clinical guideline passages from WHO IMCI danger-sign criteria [3] and Indian MoHFW PHC triage protocols [4]. The top-3 scoring passages per query are prepended to the Stage 5 prompt, providing explicit clinical evidence for high-acuity detection without requiring model fine-tuning.

4) Stages 5 and 6: Retrieval-Augmented Triage Classification: The LLM receives the normalised patient query, retrieved guideline passages, and triage label definitions. Temperature is set to 0.0 for determinism; output is a JSON object containing the triage label, a one-sentence clinical justification, and a confidence flag. Based on benchmarking results, Sarvam-105B with retrieval augmentation is recommended as the sovereign pipeline configuration.

5) Stage 7: Doctor Alert Interface: CRITICAL outputs trigger a structured immediate alert to the nearest available physician containing the audio clip, ASR transcript, retrieved passages, and LLM reasoning. CONSULT outputs schedule a telemedicine appointment within 24 hours. ROUTINE outputs direct the patient to pharmacy or self-care guidance.

B. Research Questions

RQ1. Does Saaras V3 outperform Whisper Large v3 on Hindi medical speech recognition (WER, CER)?
RQ2. On zero-shot code-mixed Hindi medical triage, does Sarvam-105B reach comparable Macro F1 to global LLMs?
RQ3. Does retrieval augmentation from WHO IMCI and Indian PHC guidelines improve triage performance, and does this effect differ between sovereign and global models?

C. Dataset 1: EkaCare Hindi Medical ASR Benchmark

The EkaCare Medical ASR Evaluation Dataset [11] contains medical speech recordings with expert-verified reference transcripts across three speech styles. We evaluated on the complete Hindi-language subset of **320 recordings at 16 kHz**. All 320 samples yielded valid predictions from both ASR systems.

D. Dataset 2: MMCQSD Triage Benchmark

The Multimodal Medical Code-Mixed Question Summarisation Dataset (MMCQSD) [17] contains 3,015 real-world Hindi-English code-mixed patient queries. We constructed a stratified sample of **500 queries** for triage benchmarking.

TABLE I
REPRESENTATIVE WEAK-LABEL ANNOTATION EXAMPLES (MMCQSD)

Patient Query (Code-Mixed)	Label	LF Trigger
"Bahut tez chest pain ho raha hai, saans nahi aa rahi, left arm mein dard"	CRITICAL	LF-A: chest pain, respiratory distress. LF-B: bahut tez, saans nahi. Three-way agreement.
"Main football khel raha tha aur bhayanak tarike se gira, sar mein chot"	CRITICAL	LF-A: head injury. LF-B: emergency language. Two-of-three agreement.
"Mere 2.8 saal ki beti ko nephrotic syndrome hua, antibiotic chal rahi hai"	CONSULT	No IMCI danger signs; persistent condition requiring physician follow-up.
"Mujhe 3 din se bukhar hai, 101 degree, saath mein khansi bhi hai"	CONSULT	Fever > 48h with cough: PHC evaluation. No danger signs.
"Kya main vitamin C supplement daily le sakta hoon?"	ROUTINE	General health query. No symptoms. Three-way agreement.

From 2,000 randomly selected queries (seed = 42), three deterministic labelling functions (LFs) were applied. The final benchmark contains CRITICAL = 80 (16.0%), CONSULT = 340 (68.0%), ROUTINE = 80 (16.0%). The 16% CRITICAL rate reflects deliberate oversampling to obtain reliable per-class statistics.

1) Weak-Label Annotation: All rules are grounded in WHO IMCI danger-sign criteria [3] and Indian MoHFW PHC triage protocols [4]. **LF-A (Summary-Level Severity Keywords).** CRITICAL if any WHO IMCI danger-sign term appears in the physician summary. ROUTINE if only self-limiting descriptors appear. All other cases: CONSULT.
LF-B (Patient Query Urgency Signals). CRITICAL if Hindi or English urgency language is present. ROUTINE if chronic presentation (> 10 days) with no deterioration indicators. All other cases: CONSULT.
LF-C (Demographic Risk Features). CRITICAL if paediatric (< 5 years) with concurrent fever and respiratory symptoms, or elderly (> 65 years) with cardiac or neurological descriptors. ROUTINE if both LF-A and LF-B yield ROUTINE; otherwise CONSULT.

Final labels were assigned by majority vote. Three-way LF agreement was obtained for 287 of 500 samples (57.4%); the remaining 213 were resolved by two-of-three majority. These labels are designated as **weak labels** produced by deterministic LFs, following established programmatic labelling practice [22].

E. Models Evaluated

Table II lists all eight models evaluated in the triage benchmark.

F. Evaluation Protocol

For ASR, both models received identical 16 kHz float32 normalised audio with language set to Hindi. WER and CER were computed against EkaCare expert-verified reference transcripts. For Triage, all LLMs received identical prompts at temperature = 0.0. Two variants were used per model: **zero-shot** (ZS) and **retrieval-augmented generation** (RAG).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

TABLE II
MODELS EVALUATED IN THIS STUDY

Model	Role	Params	Access	Provider
Saaras V3	ASR (Sovereign)	MoE	API	Sarvam AI
Whisper Large v3	ASR (Global)	1.5B	Local (CPU)	OpenAI (OSS)
Sarvam-105B	LLM (Sovereign)	105B	API	Sarvam AI
Sarvam-30B	LLM (Sovereign)	30B	API	Sarvam AI
LLaMA-3.3-70B	LLM (Global)	70B	Nebius API	Meta (OSS)
DeepSeek-V3.2	LLM (Global)	MoE	Nebius API	DeepSeek (OSS)
Qwen3-32B	LLM (Global)	32B	Nebius API	Alibaba (OSS)
GPT-OSS-20B	LLM (Global)	20B	Nebius API	OpenAI
GPT-OSS-120B	LLM (Global)	120B	Nebius API	OpenAI
Gemma-3-27B	LLM (Global)	27B	Nebius API	Google (OSS)

TABLE III
ASR RESULTS. PUBLISHED ENGLISH BASELINES FROM EKACARE [11] ARE SEPARATED FROM OUR HINDI EVALUATION; CROSS-ROW COMPARISON IS NOT VALID.

Model	WER	CER	Source	Language
Parrotlet-a-en-5b	0.109	0.047	EkaCare benchmark	English
Whisper Large v3	0.157	0.056	EkaCare benchmark	English
Bhashini ASR	0.199	0.093	EkaCare benchmark	English
Saaras V3	**0.285**	**0.122**	This work	Hindi
Whisper Large v3	0.565	0.360	This work	Hindi

TABLE IV
ASR WER BY CLIP DURATION (EKACARE HINDI SUBSET, n = 320)

Duration	n	Saaras V3	Whisper
0 to 5 s	6	0.743	1.486
5 to 10 s	6	0.392	0.409
10 to 20 s	198	0.261	0.577
20 to 30 s	110	0.304	0.543

Output was parsed as JSON. Metrics: Accuracy, per-class F1, Macro F1, CRITICAL Recall, CRITICAL Precision.

IV. RESULTS AND FINDINGS

A. ASR Performance (RQ1)

Saaras V3 achieves WER = 0.285 and CER = 0.122 (320 samples, 0 failures), compared to WER = 0.565 and CER = 0.360 for Whisper Large v3: a **49.6%** relative WER reduction and 66.1% relative CER reduction, directly supporting RQ1. Table IV breaks WER by clip duration. Saaras V3 outperforms Whisper at every duration; the largest advantage occurs at 10 to 20 second clips (n = 198, 61.9% of data; WER 0.261 vs. 0.577).

B. Zero-Shot Triage Classification (RQ2)

Table V reports zero-shot results for all eight models. GPT-OSS-120B leads at Macro F1 = 0.481. Pairwise comparison of bootstrap 95% CIs shows complete overlap for all 28 model pairs. **No model achieves statistically significantly higher Macro F1 than any other.** RQ2 is partially supported: Sarvam-105B reaches performance comparable to global LLMs (no significant gap).

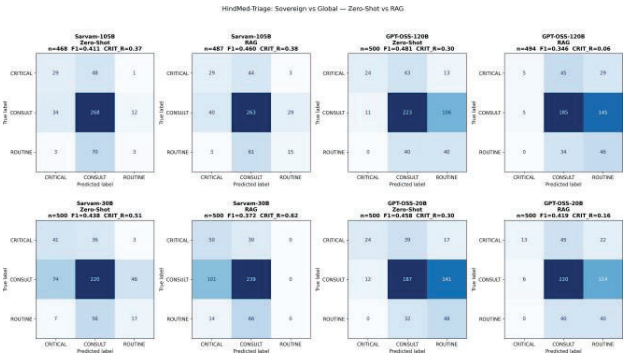

Fig. 2. Confusion matrices for all eight LLM configurations under zero-shot (ZS) and retrieval-augmented (RAG) conditions (n = 500; CRITICAL n = 80). Sarvam-30B RAG achieves the highest CRITICAL recall (0.62) with zero CRITICAL-to-ROUTINE errors. GPT-OSS-120B RAG degrades to CRITICAL recall = 0.06 with 29/80 CRITICAL cases routed to ROUTINE.

C. Retrieval-Augmented Generation Results (RQ3)

Table VI presents the complete RAG versus zero-shot comparison. **Sarvam-105B is the only model among eight for which RAG improves Macro F1** (+0.049, from 0.411 to 0.460). Every other model degrades under RAG. A notable exception to the Macro F1 trend is Sarvam-30B, whose CRITICAL recall *increases* under RAG (0.512 to 0.625). The GPT-OSS-120B RAG collapse is the most clinically significant finding. Under zero-shot, GPT-OSS-120B sends 13 of 80 CRITICAL samples to ROUTINE; under RAG, 29 of 80 are sent to ROUTINE.

D. CRITICAL Class Safety Analysis

Table VII reveals four patterns. *Sovereign models have the highest CRITICAL recall. Both GPT-OSS models exhibit a systematic CRITICAL-to-ROUTINE failure. Sarvam-30B RAG produces zero CRITICAL-to-ROUTINE misclassifications. RAG amplifies GPT-OSS danger; it suppresses it for Sarvam.*

V. DISCUSSION

A. Pipeline Design Justification

Benchmarking provides direct empirical support for two HindMed-Triage design choices. The Saaras V3 front-end is justified by the 49.6% WER advantage over Whisper on Hindi medical speech. At WER = 0.565, more than half of word tokens are transcribed incorrectly on average, generating noise that propagates through retrieval and reasoning. The retrieval augmentation stage is supported, but with model-specificity that carries clear architectural implications.

B. Sovereign vs. Global Safety Profiles

The CRITICAL class analysis separates models into two distinct safety profiles when CRITICAL-to-ROUTINE misclassification is used as the primary safety criterion. Sarvam-105B ZS, LLaMA-3.3-70B ZS, Gemma-3-27B ZS, and Sarvam-30B RAG all keep CRITICAL-to-ROUTINE errors

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

TABLE V
ZERO-SHOT TRIAGE RESULTS ON WEAK-LABEL BENCHMARK (n = 500; CRITICAL = 80, CONSULT = 340, ROUTINE = 80). BOOTSTRAP 95% CI ON MACRO F1 (n_{BOOT} = 1,000, SEED = 42). COVERAGE < 100% INDICATES PARSE ERRORS.

Model	Cov.	Acc.	Macro F1	95% CI	CRIT F1	CRIT Rec.	CONS F1	ROUT F1
GPT-OSS-120B	100%	0.574	**0.481**	[0.425, 0.537]	0.417	0.300	0.690	0.335
GPT-OSS-20B	100%	0.518	0.458	[0.406, 0.514]	0.414	0.300	0.625	0.336
DeepSeek-V3.2	100%	0.608	0.457	[0.404, 0.508]	0.358	0.275	0.730	0.282
Sarvam-30B	100%	0.556	0.438	[0.388, 0.484]	0.406	0.512	0.675	0.233
LLaMA-3.3-70B	100%	0.676	0.433	[0.382, 0.489]	0.336	0.237	0.796	0.167
Qwen3-32B	100%	0.700	0.413	[0.367, 0.458]	0.376	0.275	0.814	0.049
Gemma-3-27B	100%	0.654	0.421	[0.368, 0.471]	0.328	0.237	0.778	0.157
Sarvam-105B	94%	0.641	0.411	[0.366, 0.456]	0.403	0.372	0.766	0.065

TABLE VI
ZERO-SHOT VS. RAG: MACRO F1 DELTA AND CRITICAL RECALL. Δ = RAG − ZS. ALL EIGHT MODELS TESTED.

Model	ZS F1	RAG F1	Δ F1	CRIT Rec. ZS / RAG
Sarvam-105B	0.411	**0.460**	+0.049	0.372 / 0.382
Sarvam-30B	0.438	0.372	−0.066	0.512 / 0.625
Qwen3-32B	0.413	0.390	−0.023	0.275 / 0.250
Gemma-3-27B	0.421	0.382	−0.039	0.237 / 0.163
GPT-OSS-20B	0.458	0.419	−0.039	0.300 / 0.163
LLaMA-3.3-70B	0.433	0.416	−0.017	0.237 / 0.200
DeepSeek-V3.2	0.457	0.442	−0.015	0.275 / 0.312
GPT-OSS-120B	0.481	0.346	−0.135	0.300 / 0.063

TABLE VII
CRITICAL CLASS MISCLASSIFICATION ($n_{CRITICAL}$ = 80). C→R IS THE MOST DANGEROUS ERROR (EMERGENCY ROUTED TO SELF-CARE).

Model	Mode	C→C	C→Con	C→R	Err	Recall
Sarvam-30B	RAG	50/80	30/80	0/80	0	**0.62**
Sarvam-30B	ZS	41/80	36/80	3/80	0	0.51
Sarvam-105B	ZS	29/80	48/80	1/80	2	0.36
Sarvam-105B	RAG	29/80	44/80	3/80	4	0.36
DeepSeek-V3.2	ZS	22/80	55/80	3/80	0	0.28
DeepSeek-V3.2	RAG	25/80	52/80	3/80	0	0.31
Qwen3-32B	ZS	22/80	58/80	0/80	0	0.28
Qwen3-32B	RAG	20/80	60/80	0/80	0	0.25
GPT-OSS-20B	ZS	24/80	39/80	17/80	0	0.30
GPT-OSS-20B	RAG	13/80	45/80	22/80	0	0.16
GPT-OSS-120B	ZS	24/80	43/80	13/80	0	0.30
GPT-OSS-120B	RAG	5/80	45/80	29/80	1	0.06
LLaMA-3.3-70B	ZS	19/80	60/80	1/80	0	0.24
LLaMA-3.3-70B	RAG	16/80	63/80	1/80	0	0.20
Gemma-3-27B	ZS	19/80	60/80	1/80	0	0.24
Gemma-3-27B	RAG	13/80	67/80	0/80	0	0.16

at one or fewer. GPT-OSS-20B and GPT-OSS-120B form a distinct high-risk group. The implication is that recall alone is insufficient as a clinical safety metric; CRITICAL-to-ROUTINE error count is more directly relevant for deployment decisions.

C. Hybrid Architecture Rationale

A purely sovereign pipeline (Saaras V3 and Sarvam-105B with RAG) is the recommended configuration. It achieves the best-documented safety profile, is the only configuration that benefits from retrieval augmentation on both Macro F1 and CRITICAL recall dimensions, and avoids the 49.6% ASR degradation that any Whisper-based pipeline incurs.

D. Overlapping Confidence Intervals as an Informative Result

The complete overlap of all 28 pairwise 95% CIs indicates that LLM choice alone does not produce statistically distinguishable triage performance on this benchmark. This is an informative negative result with practical consequences: selecting a model on Macro F1 alone is not warranted.

E. Limitations

- **Weakly supervised labels.** Benchmark labels are generated via deterministic labelling functions; while suitable for relative comparisons, future work will incorporate physician-adjudicated ground truth.
- **Domain-limited ASR evaluation.** The EkaCare dataset covers a single medical domain (`misc_medical`); broader evaluation across speech styles remains future work.
- **System-level evaluation gaps.** Latency, cost, and end-to-end pipeline performance under real-world deployment conditions are not evaluated.
- **Model-specific behaviour under RAG.** The observed GPT-OSS behaviour under retrieval augmentation warrants further investigation with alternative prompt designs.

VI. CONCLUSION

We presented HindMed-Triage, a seven-stage hybrid sovereign-global AI pipeline for code-mixed Hindi medical speech triage in rural Indian telemedicine. Benchmarking across 320 Hindi ASR samples and a 500-sample stratified weak-label triage benchmark spanning eight LLM configurations yielded five findings. First, Saaras V3 reduces WER by 49.6% over Whisper Large v3 on Hindi medical speech. Second, on zero-shot triage, all eight models fall within Macro F1 of 0.411 to 0.481 with fully overlapping bootstrap

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

95% CIs. Third, retrieval augmentation improves Sarvam-105B by +0.049 Macro F1 while reducing Macro F1 for all seven other models. Fourth, sovereign models achieve higher CRITICAL recall than global models. Fifth, both GPT-OSS model variants exhibit a safety-critical pattern under zero-shot and RAG conditions misclassifying CRITICAL cases to ROUTINE.

VII. Future Scope

Future work will conduct an end-to-end pipeline evaluation including text normalisation and retrieval, develop a purpose-built triage corpus with physician adjudication and balanced class sampling, investigate the GPT-OSS RAG misapplication mechanism, and conduct field trials at Primary Health Centre facilities.

References

[1] World Health Organization, "Health Workforce," *WHO Global Health Observatory*, 2023. [Online]. Available: https://www.who.int/data/gho
[2] Ministry of Health and Family Welfare, Government of India, "Rural Health Statistics 2021," 2021.
[3] World Health Organization, "Integrated Management of Childhood Illness (IMCI): Chart Booklet," WHO/FCH/CAH/00.12, Geneva, 2005.
[4] Ministry of Health and Family Welfare, Government of India, "Indian Public Health Standards (IPHS) Guidelines for Primary Health Centres," 2019.
[5] K. Bali et al., "I am borrowing ya mixing? An analysis of English-Hindi code mixing in Facebook," in *Proc. Workshop on Computational Approaches to Code Switching (EMNLP)*, 2014, pp. 116–126.
[6] S. Khanuja et al., "MuRIL: Multilingual Representations for Indian Languages," *arXiv preprint arXiv:2103.10730*, 2021.
[7] V. Aggarwal et al., "Code-Mixed Clinical NLP: Challenges and Approaches for Hindi-English Medical Text," in *Proc. FIRE*, 2022.
[8] S. Sitaram et al., "A Survey of Code-Switched Speech and Language Processing," *arXiv preprint arXiv:1904.00784*, 2019.
[9] A. Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proc. ICML*, 2023.
[10] A. Mani et al., "Towards Automatic Speech Recognition for Medical Domain in Low-Resource Settings," in *Proc. Interspeech*, 2020.
[11] EkaCare, "Eka Medical ASR Evaluation Dataset," MIT Licence, HuggingFace: `ekacare/eka-medical-asr-evaluation-dataset`, 2024. [Online]. Available: https://huggingface.co/datasets/ekacare/eka-medical-asr-evaluation-dataset
[12] E. Salesky et al., "Multilingual Speech Translation with Efficient Finetuning of Pretrained Models," in *Proc. ACL-IJCNLP*, 2021.
[13] H. Nori et al., "Capabilities of GPT-4 on Medical Challenge Problems," *arXiv preprint arXiv:2303.13375*, 2023.
[14] T. Panch et al., "Artificial Intelligence: Opportunities and Risks for Public Health," *The Lancet Digital Health*, vol. 1, no. 1, pp. e13–e14, 2019.
[15] D. Jin et al., "What Disease Does This Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams," *Applied Sciences*, vol. 11, no. 14, p. 6421, 2021.
[16] N. Faujdar and S. Ghosh, "Analysis and Prediction of Disease Using Machine Learning Approach in Healthcare," in *Proc. Int. Conf. Computational Intelligence and Data Science*, 2022.
[17] A. Acharya et al., "MedSumm: A Multimodal Approach to Summarizing Code-Mixed Hindi-English Clinical Queries," in *Proc. European Conf. Information Retrieval (ECIR)*, 2024.
[18] Sarvam AI, "Sarvam-105B: India's Sovereign Large Language Model," Technical Report, Feb. 2026. [Online]. Available: https://www.sarvam.ai
[19] D. Kakwani et al., "IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages," in *Findings of EMNLP*, 2020.
[20] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.
[21] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.
[22] A. Ratner et al., "Data Programming: Creating Large Training Sets, Quickly," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.
[23] Meta AI, "The LLaMA 3 Herd of Models," *arXiv preprint arXiv:2407.21783*, 2024.
[24] C. Beleites et al., "Sample Size Planning for Classification Models," *Analytica Chimica Acta*, vol. 760, pp. 25–33, 2013.