A Survey Paper on Local Multilingual Medical Consultation Summarizer (LLM & Speech)

T.G. Racida[1], B.S. Prakash[2]
[1]*PG Scholar, Department of Computer Science and Engineering, Vivekanandha College of Engineering for Women, Tamilnadu, India*
[2]*Assistant Professor, Department of Computer Science and Engineering, Vivekanandha College of Engineering for Women, Tamilnadu, India*

***Abstract*—Healthcare consultations contain critical information such as symptoms, diagnosis, medicines, tests and follow-up instructions, but this information is difficult to document completely, particularly when conversations take place in local languages. Patients often forget medical advice after the visit, and manual documentation increases the workload of clinicians. This paper surveys ten core studies on bilingual and clinical speech recognition, real-time voice translation, large language model (LLM) based summarization of clinical text and medical dialogue, and privacy-preserving machine learning for e-health, together with ten supporting works on datasets, evaluation, clinical safety and earlier surveys. The core studies are compared by input, technique and limitation, and by whether they address speech input, multilingual support, medical summarization, patient-oriented output, factual consistency and local privacy. The survey shows that existing work addresses individual stages of the problem, but none integrates speech recognition, medical information extraction, factual summarization and local-language patient-friendly output in a single privacy-aware pipeline. Based on this gap, a local multilingual consultation summarizer combining multilingual Automatic Speech Recognition (ASR), speaker identification, medical entity extraction, an LLM summarizer, a factuality check and language adaptation is outlined, and open challenges are discussed.**

***Index Terms*—Medical Consultation Summarization, Automatic Speech Recognition, Large Language Models, Multilingual NLP, Patient-Friendly Summary, Privacy-Preserving AI.**

I. INTRODUCTION

A medical consultation is a short but information-rich conversation in which a doctor collects the patient's complaints, reaches a clinical impression and explains the treatment plan. The essential content of this exchange must be documented accurately, yet documentation is usually performed manually under time pressure, which leads to incomplete or inconsistent records and adds to clinician workload.

The difficulty is greater in multilingual and rural settings, where patients describe their illness in a local language or a mixture of local language and English, while clinical records are written in English. Many such conversations are never documented, and patients may forget medication instructions and follow-up advice once they leave the clinic.

Advances in Automatic Speech Recognition (ASR), Natural Language Processing (NLP) and Large Language Models (LLMs) make it possible to convert spoken consultations into text, extract the medically relevant facts and generate concise summaries. Open multilingual speech models [11] and locally runnable LLMs make on-device processing realistic, which is important because consultation audio is highly sensitive health data. However, existing research tends to treat speech recognition, translation, text-based summarization and privacy as separate problems. This paper surveys ten core studies across these areas and ten supporting works on datasets, evaluation and safety, compares their methods and limitations, and identifies the gap for a local, multilingual and patient-oriented medical consultation summarizer that takes speech as input and produces summaries in the patient's preferred language.

II. LITERATURE REVIEW

A. Speech Recognition for Multilingual and Clinical Conversations The surveyed studies on speech and translation show both the promise and the limits of

current technology for clinical conversations. Collectively, they emphasise language mixing, platform dependence and offline use, which are all central to a local multilingual summarizer.

[1] Ahmad et al. present a review, taxonomy and open challenges of bilingual ASR. The review covers multilingual modelling and approaches for recognising speech that contains two languages, and identifies code-switching and scarce annotated data as open problems. As a review, it does not offer a complete medical consultation summarizer.

[2] Zeeshan, Bogue and Mamoona compare diverse ASR platforms for transcribing psychiatric treatment sessions, evaluating how well clinician–patient speech is transcribed. The findings show that platform choice strongly affects transcript quality. The study is limited to psychiatric sessions and does not address multilingual medical summarization.

[3] Kanchanadevi et al. propose real-time bidirectional voice translation for healthcare using speech recognition and NLP, with offline support. The system helps doctors and patients who do not share a language, but it focuses on translation and communication rather than summarizing the consultation.

B. LLM-Based Summarization of Clinical Text and Medical Dialogue

A second group of studies concentrates on generating summaries from clinical text or text dialogues. These works show that explicitly modelling medical key information and factual consistency improves summary quality.

[4] Verma, Darlan and Mallipeddi apply multi-objective optimization to LLM-generated summaries of clinical text narratives, balancing objectives such as clinical readability. The approach is applied to written narratives and is not an end-to-end multilingual speech pipeline.

[5] Liu et al. introduce MediExtract, a distillation framework in which an LLM generates medical key information as a teacher for a smaller student summarizer, evaluated on the MTS-Dialogue dataset. It requires structured training data and works on text dialogue rather than speech.

[6] Lu, Duan and Liu present a factual-aware two-stage model for medical dialogue summarization that stresses factuality and medical terminology matching. It directly tackles unsupported content in summaries, but accepts text only and does not consider local-language speech.

[7] Duan, Lu and Liu propose a compacting-abstractive model for efficient medical dialogue summarization. Efficiency is valuable for local deployment, but multilingual ASR and patient-language generation fall outside the scope of the work.

[8] Neupane et al. present ClinicSum, which generates SOAP-style clinical summaries from patient–doctor conversations using retrieval-based filtering with fine-tuned pre-trained language models and retrieval-augmented generation. It depends on transcript text and gives limited attention to multilingual local deployment.

[9] Santos et al. evaluate six LLMs for diagnosis summarization using F1-based assessment. The study indicates which models are able to capture diagnostic information, but its focus is on diagnosis or discharge summaries and it is neither speech-based nor multilingual.

C. Related Surveys and Evidence for LLM Summarization

[15] Van Veen et al. adapted eight LLMs to four clinical summarization tasks, including doctor–patient dialogue, and in a reader study with ten physicians the best summaries were judged equivalent (45%) or superior (36%) to those of medical experts in most cases. These supports using LLMs to reduce documentation burden, although the accompanying safety analysis shows that fabricated information remains a risk.

[16] Feng et al. give a survey of dialogue summarization covering meeting, chat, email, customer-service and medical dialogue, with public datasets and unified-metric leaderboards. They identify faithfulness and multilingual summarization as future directions.

[17] Shi et al. survey medical dialogue systems by categories, methods, evaluation and challenges, and discuss the effect of LLMs. Palanisamy et al. [18] survey GPT-based summarization of medical and biomedical artifacts, including medical dialogues. Neither addresses a speech-to-local-language pipeline.

D. Datasets, Evaluation and Clinical Safety

[13] Ben Abacha et al. introduce MTS-Dialog, a collection of about 1,700 doctor–patient dialogues with clinical notes, and evaluate note generation with

n-gram, embedding-based and fact-extraction metrics supported by expert review. Yim et al. [14] provide ACI-Bench, a benchmark of doctor–patient visit dialogues paired with full clinical notes.

[19] Asgari et al. propose a framework to assess the clinical safety and hallucination rates of LLMs in medical text summarisation, with an error taxonomy and a harm framework applied to clinician-annotated sentences. Croxford et al. [20] validate an LLM-as-a-judge method for evaluating clinical summaries against human review. Both stress that automatic and human evaluation must be combined.

E. Privacy-Preserving Machine Learning for E-Health

[10] Romeo, Mahmoud, Sheriff et al. survey privacy-preserving machine learning for e-health, including differential privacy, secure computation and encryption. The survey supports keeping sensitive health data local, but as survey-level work it does not implement an end-to-end consultation summarizer.

III. COMPARATIVE ANALYSIS

The surveyed works address different tasks and use different datasets and metrics, so a single numerical performance comparison would not be meaningful. Table I therefore compares the studies by input, technique and limitation, and Table II records which aspects of the target problem each study addresses, based on its stated scope.

Table I. Summary of surveyed works

Ref	Year	Input / Task	Technique	Limitation
[1]	2022	Bilingual speech recognition (review)	Taxonomy; multilingual modelling	No complete medical consultation summarizer
[2]	2025	Clinician–patient speech transcription	Comparison of ASR platforms	Psychiatric sessions only; no multilingual summarization
[3]	2024	Speech translation in healthcare	ASR + NLP; bidirectional translation; offline	Translation focus, not consultation summarization
[4]	2026	Clinical text narratives	LLM summaries; multi-objective optimization	Text only; not an end-to-end speech pipeline
[5]	2024	Medical dialogue (text)	LLM key information; teacher–student distillation	Needs structured training data; text dialogue only
[6]	2023	Medical dialogue (text)	Two-stage; factuality; terminology matching	Text-only; no local multilingual speech
[7]	2023	Medical dialogue (text)	Compact abstractive summarization	Efficiency-focused; no ASR or patient language
[8]	2024	Patient–doctor conversation transcripts	SOAP; retrieval filtering; PLM; RAG	Transcript-dependent; limited multilingual focus
[9]	2024	Diagnosis summarization	Evaluation of six LLMs; F1 assessment	Not speech-based or multilingual
[10]	2024	E-health data privacy	Differential privacy; secure computation; encryption	Survey only; no implementation

Table II. Coverage of key aspects in the surveyed works

Ref	Speech / ASR	Multilingual	Medical Summary	Patient-Oriented	Factuality	Local Privacy
[1]	Yes	Yes	–	–	–	–
[2]	Yes	–	–	–	–	–
[3]	Yes	Yes	–	Partial	–	Partial
[4]	–	–	Yes	Partial	–	–
[5]	–	–	Yes	–	Partial	–
[6]	–	–	Yes	–	Yes	–
[7]	–	–	Yes	–	–	–
[8]	–	–	Yes	–	Partial	–
[9]	–	–	Yes	–	–	–
[10]	–	–	–	–	–	Yes
Proposed	Yes	Yes	Yes	Yes	Yes	Yes

Yes = addressed as a main focus; Partial = addressed in part; – = not addressed in the work's stated scope.

Table III. Public datasets and models relevant to the project

Resource	Type	Relevance to the Project
MTS-Dialog [13]	Doctor–patient dialogue and notes (about 1,700 pairs)	Training and testing of dialogue-to-note summarization
ACI-Bench [14]	Visit dialogues with full clinical notes	Benchmark for full-note generation from conversation
Whisper [11]	Multilingual ASR model (680,000 hours of training audio)	Candidate for local multilingual speech recognition
IndicTrans2 [12]	Translation models for 22 Indian languages	Candidate for Tamil patient-language adaptation

IV. PROPOSED FRAMEWORK

Based on the gaps identified in Tables I and II, a local multilingual medical consultation summarizer is outlined in Fig. 1. The consultation audio is captured locally and converted into a transcript by a multilingual ASR model, for which Whisper-style models [11] are candidates. Speakers are identified and the transcript is cleaned of noise, repetition and irrelevant content. A medical NLP stage extracts symptoms, diagnosis, medicines, tests and follow-up instructions. A medical-domain LLM then generates a concise, factual summary, a factuality and safety check compares it with the transcript, and a language-adaptation stage translates and simplifies it for the patient, for example in Tamil, using models such as IndicTrans2 [12]. Wherever feasible, audio, transcripts and inference stay on the local machine. To ensure strict privacy and low-latency execution in outpatient clinics, the proposed pipeline employs quantized local foundation models (e.g., Faster-Whisper Medium using CTranslate2 for acoustic transcription and 4-bit GGUF quantized medical LLMs such as BioMistral-7B). This restricts memory consumption to under 10 GB of VRAM, making edge deployment feasible on standard medical workstation hardware without active internet connectivity.

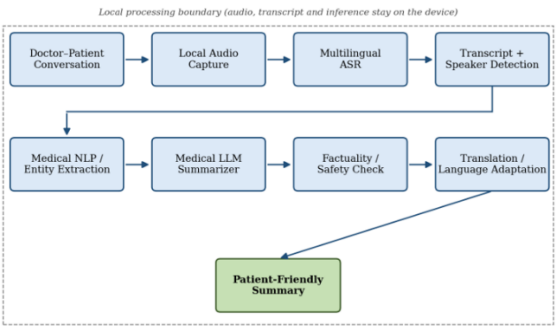

Fig. 1. Proposed system architecture of the local multilingual medical consultation summarizer

V. ANALYSIS

The survey shows that each stage of consultation summarization has been studied, but mostly in isolation. Speech-oriented studies [1]–[3] deal with recognition and translation, yet stop at the transcript or the translated utterance. Summarization studies [4]–[9] produce good summaries from text, but assume a clean transcript and are written for clinicians rather than patients. Privacy work [10] stresses protecting health data but offers no consultation pipeline.

Several themes recur. First, factual consistency is essential: approaches that extract medical key information or enforce terminology matching [5], [6] improve reliability and motivate a dedicated factuality check. Second, structured output such as SOAP notes combined with retrieval [8] gives controllable summaries. Third, ASR quality depends strongly on the platform and the type of clinical speech [2], so recognition errors in drug names and dosages can propagate into the summary. Fourth, efficiency [7] and offline operation [3] indicate that local deployment is feasible.

Open challenges remain. Code-mixed local-language speech is poorly supported [1]; patient-friendly output in the patient's language is rarely evaluated; and hallucination in LLM summaries is a safety risk that needs both automatic checks and clinician review. Existing datasets such as MTS-Dialog [13] and ACI-Bench [14] are English text dialogues, so no public resource pairs local-language consultation audio with notes and patient summaries; creating such data is itself a research need. Evaluation should combine automatic metrics with clinician review and safety analysis [19], [20]. A complete solution should therefore keep the doctor in the loop and be evaluated with word error rate, ROUGE, BERTScore, entity-level F1, hallucination rate and clinician ratings.

Future directions include collecting consented Tamil and code-mixed consultation recordings, fine-tuning ASR and summarization models on them, adding text-to-speech output for patients with limited literacy, and validating the pipeline in a clinical usability study.

VI. CONCLUSION

This paper surveyed ten core studies and ten supporting works relevant to a local multilingual medical consultation summarizer. Speech recognition, voice translation, LLM-based medical summarization and privacy-preserving learning have each advanced, but no surveyed work combines speech input, medical information extraction, factual summarization, local-language patient-friendly output and local processing. A pipeline combining multilingual ASR, speaker identification, medical entity extraction, an LLM summarizer, a factuality check and language adaptation is outlined to address this gap. Future work will implement and evaluate the pipeline on English, Tamil and code-mixed consultations, with doctor review of the generated summaries.

REFERENCES

[1]. Abushariah, A. A. M., Ting, H.-N., Mustafa, M. B. P., Khairuddin, A. S. M., Abushariah, M. A. M., and Tan, T.-P., "Bilingual automatic speech recognition: A review, taxonomy and open challenges," IEEE Access, vol. 11, pp. 5944–5954, 2023, doi: 10.1109/ACCESS.2022.3218684.

[2]. Zeeshan, R., Bogue, J., and Malik, M. K., "Relative applicability of diverse ASR platforms for transcription of psychiatric treatment sessions," in Proceedings of the International Conference on Health Informatics and Medical Systems (HIMS), Springer, 2025, pp. 45–58.

[3]. Kanchanadevi, P., Illakkia, A., Janane, S., and Dheepika, I., "Enhancing patient care through NLP-based realtime bidirectional voice translation in healthcare," in Proceedings of the IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), 2024, pp. 1–6, doi: 10.1109/ICSCAN61559.2024.10894641.

[4]. Verma, D., Darlan, D., and Mallipeddi, R., "Multi-objective optimization of large language model summaries for clinical text narratives," IEEE Access, vol. 14, pp. 13597–13609, 2026, doi: 10.1109/ACCESS.2026.3533214.

[5]. Liu, X., Huang, M., Rusnachenko, N., Ive, J., Chang, J., and Zhang, J. J., "Enhancing medical dialogue summarization: A MediExtract distillation framework," in Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2024, pp. 3120–3127.

[6]. Lu, F., Duan, J., and Liu, J., "A factual aware two-stage model for medical dialogue summarization," in Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2023, pp. 2859–2866, doi: 10.1109/BIBM58861.2023.10385609.

[7]. Duan, J., Lu, F., and Liu, J., "Towards efficient medical dialogue summarization with compacting-abstractive model," in Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2023, pp. 2867–2874, doi: 10.1109/BIBM58861.2023.10385437.

[8]. Neupane, S., Karunanayake, I., Lin, C., et al., "ClinicSum: Utilizing language models for generating clinical summaries from patient-doctor conversations," in Proceedings of the IEEE International Conference on Big Data (BigData Workshop on AI for Healthcare), 2024, pp. 1–8, arXiv:2412.04254.

[9]. Santos, J., Santos, H. D. P., Ulbrich, A. H. D. P. S., et al., "Evaluating LLMs for diagnosis summarization," in Proceedings of the 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2024, pp. 1–7, doi: 10.1109/EMBC53108.2024.10782341.

[10]. Romeo, J., Gohar, M., and Sheriff, A., "Privacy-preserving machine learning for e-health applications: A survey," in Proceedings of the 2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI), 2024, pp. 1–7, doi: 10.1109/ICMI61559.2024.10586115.

[11]. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I., "Robust speech recognition via large-scale weak supervision," in Proceedings of the 40th International Conference

on Machine Learning (ICML), vol. 202, PMLR, 2023, pp. 28492–28518.

[12]. Gala, J., Chitale, P. A., et al., "IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages," Transactions on Machine Learning Research, 2023, arXiv:2305.16307.

[13]. Ben Abacha, A., Yim, W., Fan, Y., and Lin, T., "An empirical study of clinical note generation from doctor-patient encounters," in Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2023, pp. 2291–2302, doi: 10.18653/v1/2023.eacl-main.168.

[14]. Yim, W., Fu, Y., Ben Abacha, A., Snider, N., Lin, T., and Yetisgen, M., "ACI-BENCH: A novel ambient clinical intelligence dataset for benchmarking automatic visit note generation," Scientific Data, vol. 10, no. 1, Art. no. 591, pp. 1–11, 2023, doi: 10.1038/s41597-023-02487-3.

[15]. Van Veen, D., Van Uden, C., Blankemeier, L., et al., "Adapted large language models can outperform medical experts in clinical text summarization," Nature Medicine, vol. 30, no. 4, pp. 1134–1142, 2024, doi: 10.1038/s41591-024-02855-5.

[16]. Feng, X., Feng, X., and Qin, B., "A survey on dialogue summarization: Recent advances and new frontiers," in Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI), 2022, pp. 5453–5460, doi: 10.24963/ijcai.2022/764.

[17]. Shi, X., Liu, Z., Du, L., et al., "Medical dialogue: A survey of categories, methods, evaluation and challenges," arXiv:2405.10630, 2024.

[18]. Palanisamy, B., Chakrabarti, A., Singh, A., Hassija, V., Chalapathi, G. S. S., and Singh, A., "From information overload to lucidity: A survey on leveraging GPTs for systematic summarization of medical and biomedical artifacts," IEEE Access, vol. 13, pp. 7902–7922, 2025, doi: 10.1109/ACCESS.2024.3521596.

[19]. Asgari, E., Montaña-Brown, N., Dubois, M., et al., "A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation," npj Digital Medicine, vol. 8, Art. no. 274, pp. 1–12, 2025, doi: 10.1038/s41746-025-01824-7.

[20]. Croxford, E., Gao, Y., First, E., et al., "Evaluating clinical AI summaries with large language models as judges," npj Digital Medicine, vol. 8, no. 1, Art. no. 118, pp. 1–10, 2025, doi: 10.1038/s41746-025-02005-2.