InTriage: Intelligent Telephone Triage in Pre-Hospital Emergency Care

Kai He[1], Qika Lin[1], Hao Fei[1],
Chng Eng Siong[2], Dehan Hong[3], Marcus Eng Hock Ong[*4,5], Mengling Feng[*1],

[1]National University of Singapore, Singapore
[2]Nanyang Technological University, Singapore
[3]Singapore Civil Defence Force, Singapore
[4]Duke-NUS Medical School, Singapore
[5]Singapore General Hospital, Singapore

Abstract

Pre-hospital Emergency Care (PEC) systems are critical for managing life-threatening emergencies where rapid intervention can significantly impact patient outcomes. The rising global demand for PEC services, coupled with increased emergency calls and strained emergency departments, necessitates efficient resource utilization through Telephone Triage (TT) systems. However, existing TT processes face challenges such as incomplete data collection, communication barriers, and manual errors, leading to high over-triage and under-triage rates. This study proposes InTriage, an AI-driven multilingual TT system to provide decision support for triage. InTriage enhances accuracy by transcribing emergency calls, extracting critical patient information, prompting supplementary, and providing real-time triage decisions support. We conducted an evaluation on a real-world corpus of approximately 40 hours of telephone data, achieving a word error rate of 14.57% for speech recognition and an F1 score of 73.34% for key information extraction. By improving communication efficiency and reducing triage errors, InTriage offers a scalable solution to potentially help address the growing demands on PEC systems globally[1].

1 Introduction

Pre-hospital Emergency Care (PEC) refers to medical assistance provided to patients outside hospital settings, typically followed by their transfer to the nearest medical facility (Mohammadi et al., 2022). PEC plays a vital role in managing life-threatening emergencies where every second is crucial, as delays can determine the outcome between survival, permanent disability, or death. Rapid patient assessment and timely intervention by PEC personnel are therefore indispensable. Globally, the demand for PEC services has been rising, as evidenced by increasing emergency calls, longer ambulance response times, and overcrowded Emergency Departments (EDs) (Brady, 2020; Inokuchi et al., 2022; Singapore Civil Defence Force, 2023). These trends highlight the strain on PEC systems and underscore the need for innovative strategies to maintain their efficiency and effectiveness.

Figure 1: Yearly data growth of emergency calls from 2000 to 2023 in Singapore. The data illustrates a quick upward trend with peaks in 2021 and 2022, reflecting significant need for more efficient triage systems.

Telephone Triage (TT) is a crucial process in PEC systems, enabling call-takers to systematically gather patient information and assess clinical severity through structured questioning. This process, which begins when an emergency call is answered and concludes with determining dispatch priority or redirecting low-acuity cases to alternative care pathways (ACPs), plays a pivotal role in optimizing PEC resource utilization. In Singapore, the number of emergency calls has grown by 400% over the past two decades, with a marked acceleration during the COVID-19 pandemic (see Figure 1). This rapid growth has posed substantial challenges to existing TT systems, which heavily rely on manual processes and call-taker expertise.

Under such conditions, a robust TT system is

[1]The introduction video is at https://www.youtube.com/watch?v=a-XnPJ4dlyw, and the demonstration video at https://www.youtube.com/watch?v=dVYonyK5-cY. The symbol (*) denotes co-corresponding authors.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 873–885
November 4-9, 2025

essential for prioritizing high-acuity, time-sensitive cases while redirecting low-acuity cases to appropriate healthcare services. This reduces the burden on EDs and ensures that critically ill patients receive prompt and focused care (Vicente et al., 2013). However, TT performance has been shown to be suboptimal in many countries. A study found an over-triage rate of 26.0% and an under-triage rate of 4.9% in Thailand, suggesting a higher prevalence of over-triage compared to under-triage (Huabbangyang et al., 2023). Another study found that over-triage rates ranged from 9.9% to 87.4%, while under-triage rates varied from 1.6% to 72.0% in North America, depending on the population and methodology (Lupton et al., 2023). In Singapore, our investigation indicates that over/under-triage rates are approximately 45% and 5%, underscoring the urgent need for improvements.

Three primary challenges contribute to these inefficiencies. First, emergency call-takers often struggle to gather essential information and assess case acuity, even with established protocols. Key data, such as medical history, clinical risk factors, and previous call records, are frequently underutilized (Wu et al., 2024; He et al., 2019). Second, communication barriers between callers and call-takers delay responses and may lead to adverse outcomes. For example, obtaining accurate location information in a multilingual country like Singapore is challenging due to diverse accents and variations names. Third, manual data entry during the triage is time-consuming and prone to human error, reducing both efficiency and accuracy.

To address these issues, we propose an artificial intelligence-based system ***InTriage***, as shown in Figure 2, which aims to assist call-takers and improve TT performance. *InTriage* leverages a multilingual Automatic Speech Recognition (ASR) agent capable of transcribing emergency calls in English, Chinese, and Malay simultaneously. Using Hotword boosting technology (Yang et al., 2024), the ASR enhances the recognition of location names. The transcriptions from the ASR agent enable a Natural Language Understanding (NLU) agent to extract critical patient information, primary complaints, and assess the caller's stress levels. The system also highlights pertinent medical history and recommends subsequent questions, ensuring a streamlined and accurate triage process. A Dialogue Management (DM) agent continuously calculates a confidence score, and once a predefined threshold is met, a Natural Language Generation (NLG) agent delivers the final triage recommendations and emergency instructions.

By automating several aspects of the triage process, *InTriage* enhances decision-making accuracy, reduces over-triage and under-triage rates, and ensures efficient use of PEC resources. Furthermore, its integration of advanced AI capabilities, addresses the critical need for precise and timely triage in high-pressure emergency settings.

To the best of our knowledge, our study presents the first intelligent multimodal TT system, which designed to meet the demands of PEC in a multilingual context. *InTriage* has been tested by the Singapore Civil Defence Force[2] (SCDF), achieving satisfactory results and offering a scalable solution for wider adoption. This work pioneers a intelligent TT systems in the following aspects.

- *InTriage* introduces a multimodal AI pipeline that seamlessly integrates five different agents. InTriage enables precise and real-time processing of emergency calls, reducing the call-taker's burden under high-pressure PEC conditions.
- *InTriage* is specifically designed to operate in multilingual environments, utilizing advanced Hotword boosting to address one of the most critical and challenging issues.
- *InTriage* integrates stress-level analysis, medical history retrieval, and emergent instructions generation, providing more comprehensive information to avoid potential risks.

2 Related Work

Different Goals and Scopes. Many previous studies on triage systems are designed for mass casualty incidents (MCIs), which are involving various approaches, including wearable technologies, electronic tagging, and telemedicine solutions. For instance, Adler et al. (2011), Greiner-Mai and Donner (2010), and Park (2021) introduced IT-supported and IoT-based e-triage systems that monitor vital signs and provide real-time updates. Besides, some triage systems are designed for disease-specific scenarios to improve diagnosis and care during outbreaks. For respiratory infections like COVID-19, Villafuerte et al. (2023) developed a

[2]The Singapore Civil Defence Force (SCDF) is a uniformed government organization under the Ministry of Home Affairs. It provides essential national services such as firefighting, rescue operations, and emergency medical services.

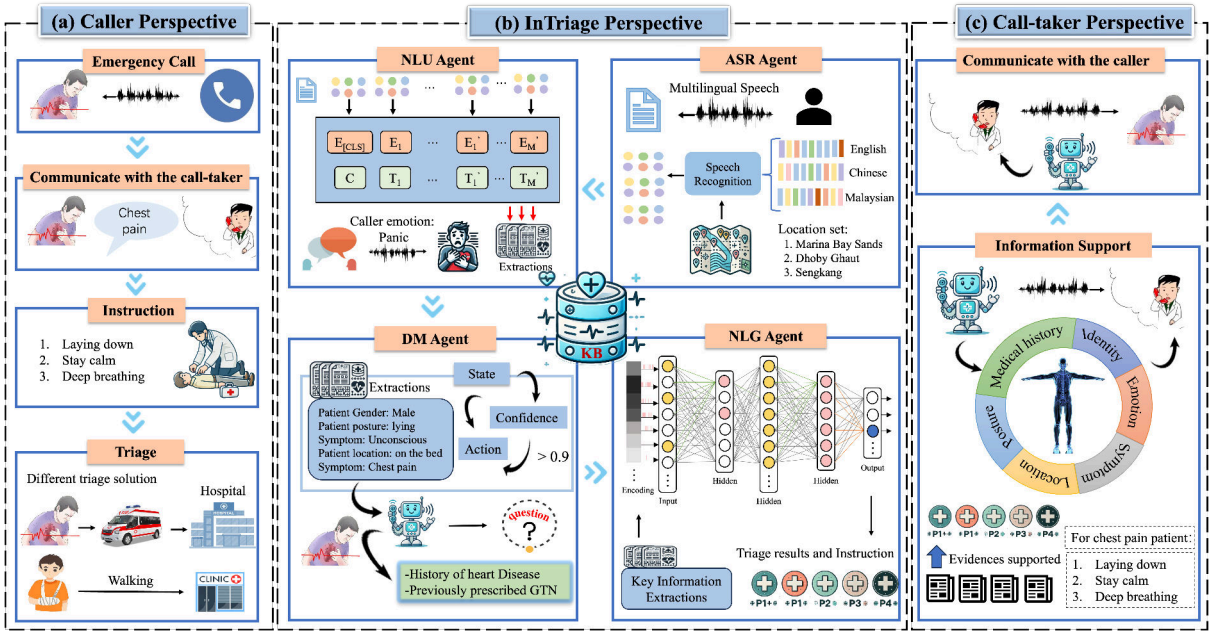

Figure 2: The architecture of InTriage, enhancing efficiency and ensuring accurate triage. We illustrate the functions of *InTriage* from three different perspectives, including the caller, InTriage, and call-taker.

telemedicine virtual assistant to diagnose conditions based on real-time vital signs and symptom evaluations. Khanna et al. (2023) and Soltan et al. (2021) employed ML models to predict COVID-19 severity using clinical biomarkers and structured data for identifying critical patients. These systems have different goals and scopes compared with broad-spectrum emergency triage *InTriage*, which is designed to serve general city residents in PEC, addressing diverse emergency scenarios without additional hardware dependencies.

Different Technologies. There are many different technologies are involved (He et al., 2025; Lin et al., 2025). In CV-based approaches, Lu et al. (2023) introduced an unmanned aerial vehicle (UAV) triage system using OpenPose and YOLO, highlighting the role of visual recognition. In ML-based solutions, Elhaj et al. (2023) conducted a comprehensive comparison of nine supervised algorithms to predict EDs outcomes, with Random Forest achieving the highest performance in ICU admission predictions, while Chen et al. (2023b) employed eXtreme Gradient Boosting (XGB) for dynamic risk stratification using ECG and chest X-ray data. In DL domain, Xiao et al. (2023) proposed TransNet and TextRNN models that integrate structured and unstructured medical data using attention mechanisms, whereas Chen et al. (2023a) developed a BiLSTM-based system that leverages clinical narratives to predict critical outcomes in EDs. In contrast, our *InTriage* system utilizes LLMs to dynamically process multimodal inputs, including real-time text, speech, and medical data, enabling more precise triage results and offering auxiliary functions like sentiment analysis and medical information retrieval to enhance decision-making in diverse emergency scenarios.

3 Preliminary Data

In TT systems, call-taker performance can vary significantly based on experience and training, leading to inconsistent triage results and potentially adverse outcomes. To address this challenge, *InTriage* utilizes the Patient Acuity Category Scale (PACS) protocol standardize the assessment process, which was introduced by Singapore's Ministry of Health (Fong et al., 2018). As shown in Table 1, PACS categorizes patients into one of five urgency levels, from P1+ (the most critical) to P4 (non-urgent). This structured approach minimizes variability in call-takers' decisions.

Under the PACS protocol, call-takers first conduct a preliminary assessment by asking general questions about the patient's information and symptoms. Based on this initial information, the case is classified into one of 30 Chief Incident Types (CITs), such as BLEEDING/LACERATION, INHALATION, or CHEST PAIN. After identifying the appropriate CIT, call-takers proceed to ask CIT-specific questions to gather more detailed informa-

PACS	Definition	Response	Example
P1+	Life-threatening emergencies	Highest priority, fastest response (extra resources deployed)	Cardiac arrest
P1		Highest priority, fastest response	Head injury
P2	Emergencies	High priority, fast response	Abdominal pain
P3	Minor Emergencies	Lower priority, slower response	Persistent Diarrhea
P4	Non-emergencies	No response	Cough

Table 1: Definitions of Patient Acuity Category Scale (PACS).

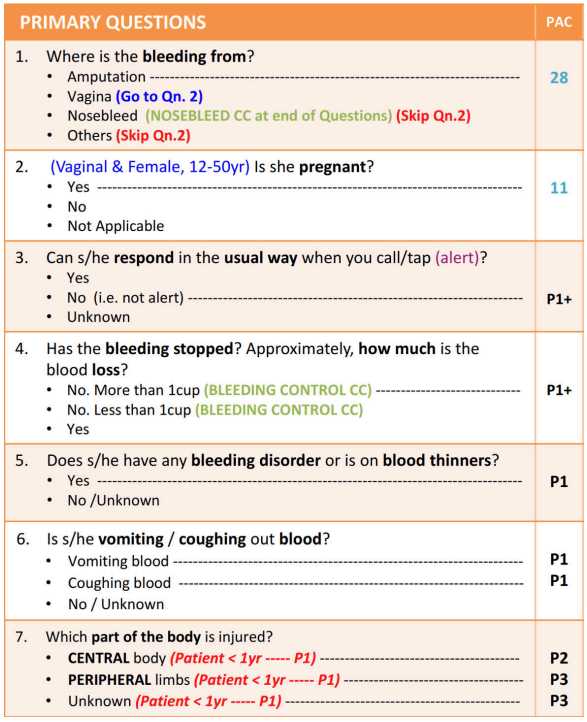

Figure 3: An example of BLEEDING/LACERATION incident type using PACS to guide call-takers for an emergency calls. The numbers in the PACS column indicate when it should be reassigned to another incident type, while P1+, P1, P2, and P3 represent triage levels.

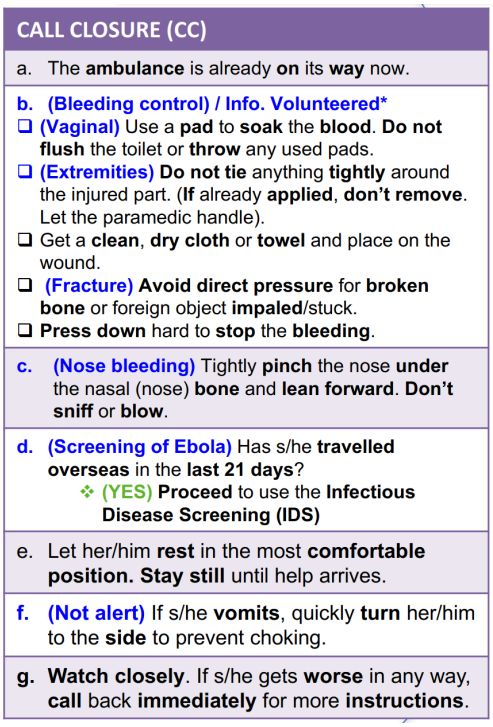

Figure 4: An example of BLEEDING/LACERATION incident type to guide call takers give necessary emergency instructions.

tion and assign the case to the appropriate PACS category for dispatch purposes. For example, in the case of BLEEDING/LACERATION, the call-taker would ask pre-defined questions such as, "Where is the bleeding from?" or "Has the bleeding stopped?" Figure 3 illustrates a sample decision tree for this specific CIT. Based on the caller's responses, the call-taker assigns the patient to a PACS category.

For assigning triage categories, call-takers must deliver CIT-specific instructions to callers to ensure the patient's safety while waiting for medical assistance. These instructions are tailored to each incident type, aiming to prevent deterioration and stabilize the patient. Figure 4 shows an example of instructions for BLEEDING/LACERATION, such as advising the caller to keep the patient still and elevate the injured area to reduce bleeding.

4 Different Perspectives from Caller, Call-taker, and InTriage

Caller Perspective – System Significance. For callers, they will no realize that any AI system is in place. As in Figure 2 (a), when they contact the emergency hotline, they are greeted by a human call-taker who responds professionally and promptly. They can also get emergency instructions, and obtain emotional comfort.

Call-taker Perspective – System Functions. For call-takers, our system provides an intelligent assistant that eliminates the need for additional manual operations to structure the content of current call. Besides, triage requires call-takers to follow strict guidelines, but the variety of emer-

gencies and question sets can overwhelm new staff and cause errors, even among experienced operators. To mitigate this issue, *InTriage* automatically recognizes the type of emergency and prompts the call-taker with the appropriate questions to ask. Additionally, our system can provide extra medical history to ensure call-takers do not overlook potential risks. Finally, *InTriage* recommends necessary emergency instructions and triage results with supporting evidence. This feature reduces the call-takers' workload and helps correct inappropriate triage decisions, ensuring better outcomes and increased efficiency in handling emergency calls.

InTriage Perspective – System Implementation. As shown in Figure 2 (b), *InTriage* consists of five agents of ASR, NLU, DM, KB, and NLG.

ASR Agent. Our ASR agent employs WeNet (Yao et al., 2021) as the backbone. We collect extra training data from audiobooks, podcasts, YouTube, and SCDF to further fine-tune WeNet, as shown in Appendix Table 3. One principle of data selection is that we have chosen audio that fits the specific accents of Singapore, including Singaporean-English, Mandarin-English, and Malay-English. At present, *InTriage* is multilingual, which can support English, Chinese, and Malay simultaneously.

NLU Agent. Our NLU agent utilizes Llama-3.2-1B. We manually annotate 2008 real cases from SCDF for training extracting key information from emergency calls. The annotated data are formatted as a QA task. For example, when an emergency call is going, the NLU agent will keep raising question like "What is the gender of the patient/caller?", "Where is the patient/caller?", or "How is the patient's mood? Describe the patient's mood and rate it on a scale of 1 to 10", NLU agent will answer the questions to automatically fill the slots.

DM Agent. Our DM agent performs three core functions. First, based on the incident type identified by the NLU agent, the DM agent provides real-time prompts to call-takers, guiding them on which questions to ask in accordance with SCDF's established protocols (see Figure 3). These prompts adjust dynamically as the NLU agent continuously refines its assessment of the incident type. Second, the DM agent retrieves data from the EHR system to obtain the caller's relevant medical history by matching extracted identity information. NLU agent further processes this data to extract key details for display. Finally, DM agent continuously calculates a confidence score. Each caller response that matches a predefined question from SCDF's protocols contributes to an incremental score.

Languages		Kaldi	Whisper	Ours	RTF
Singaporean-English	MSF Call Center	**18.25**	22.13	26.12	0.48
	MSF-DHL	26.53	24.12	**20.94**	0.42
	NTU Inhouse (1)	17.09	19.40	**10.31**	0.40
	NTU Inhouse (2)	31.60	25.85	**17.68**	0.48
	SCDF testset (1)	14.40	17.44	**10.60**	0.39
	SCDF testset (2)	28.25	24.98	**14.76**	0.46
	IMDA testset (1)	10.54	13.41	**8.60**	0.38
	IMDA testset (2)	21.95	17.43	**7.12**	0.34
	IMDA testset (2) Hotword Boosting	-	-	**4.56**	0.38
	SCDF-Mandarin	23.47	19.89	**15.29**	0.37
Malay-English	SCDF-Malay	24.16	22.65	**14.25**	0.34
	-	21.62	20.73	**14.57**	0.40

Table 2: The table compares word error rates across models. Kaldi (Povey et al., 2011) and Whisper (Radford et al., 2022) serve as baseline models. RTF (Real-Time Factor) is reported to assess our system's real-time performance on an RTX 4090 GPU.

KB Agent. KB Agent consists of an EHR database and predefined emergency instructions, which can be used for retrieval.

NLG Agent. When the score calculated by the DM agent exceeds the predefined threshold, the NLG agent generates the final triage decision and provides corresponding instructions. The NLG agent shares a common LLM with the NLU agent.

5 Evaluation and Showcases

To quantitatively evaluate the performance of the proposed *InTriage*, we conducted evaluations of both the ASR agent and NLU agent. Table 2 compares Kaldi, Whisper (small version, with larger size than ours) and our ASR agent across various cases, for Singaporean-English, Mandarin-English, and Malay-English speech recognition. All data are sampled from real-world data in Singapore. Our ASR agent outperforms baselines, showing significant improvements in word error rate, especially in Singaporean-English. Notably, the proposed model achieves a 7.12 error rate on the Imda (2) compared to Kaldi's 21.95 and further improves to 4.56 with Hotword Boosting. This result highlights the inherent challenge of accurately identifying locations in speech recognition. The model also performs better on SCDF-Mandarin and SCDF-Malay, highlighting its effectiveness in multilingual scenarios.

In Figure 6, we compare the accuracy of different caller's attributes between TANL (Paolini et al., 2021), LIama-3.1 8B, and our NLU agent. Overall, our method outperforms TANL across all attributes, with the most significant improvement

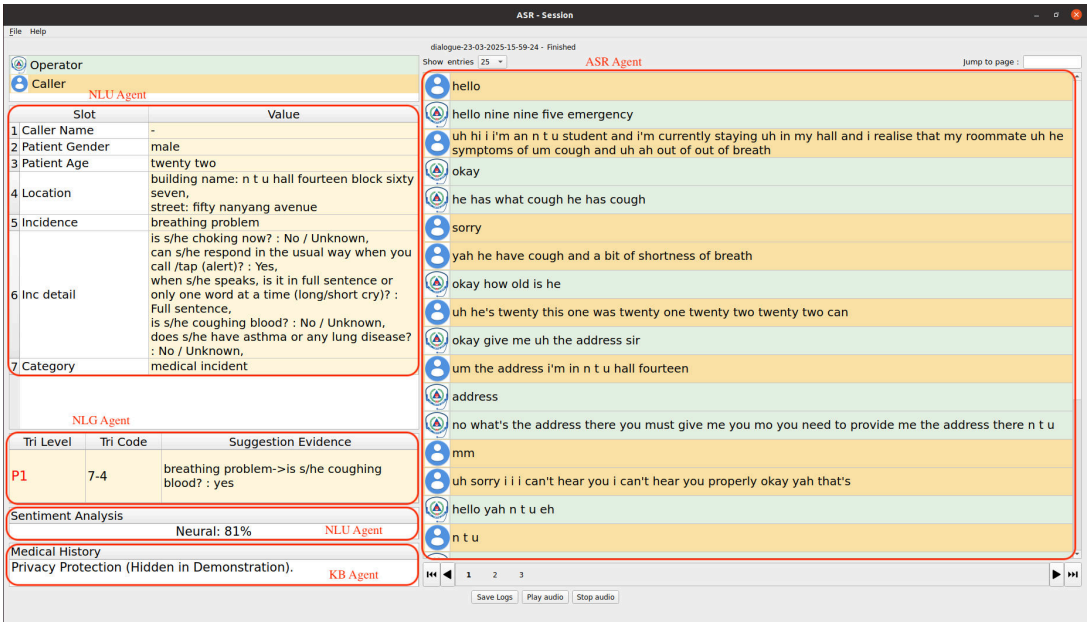

Figure 5: Screenshot of *InTriage* for an emergency case with P1 level. More show cases with different triage levels are in Appendix. All show cases in the paper use simulated conversations to ensure privacy.

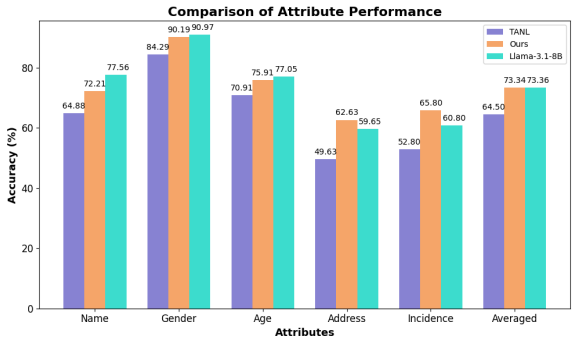

Figure 6: Breakdown analysis of key information extraction from emergency calls.

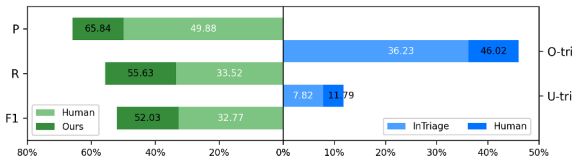

Figure 7: The triage performance comparison between InTriage with human call-taker. O-tri and U-tri indicate Over-triage and Under-triage. More detail breakdown results can be seen in appendix.

observed in the Gender attribute. The Address attribute shows the lowest performance for both methods, although our model achieves a notable improvement from 49.63% to 62.63%. Due to we take the extract matching as evaluation strategy, this performance for location recognition is satisfied. The average accuracy of our method reaches 73.34%, significantly surpassing TANL's 64.50%. This demonstrates the effectiveness of our approach across various attribute types, particularly in challenging tasks such as Incidence and Address prediction. LIama-3.1 8B performs slightly better than our NLU agent overall, but our NLU agent is only 1B in size and runs relatively faster. Considering real-time ability, our agent is more suitable when taking all factors into account. Figure 7 presents a comparison of triage performance between InTriage and human call-takers. The results show that InTriage achieves 19.25%, 22.11%, and 15.96% higher Precision, Recall, and F1-score, respectively, compared to human call-takers. Meanwhile, the rates of Over-triage and Under-triage are reduced by 9.79% and 3.97%, respectively. Figure 5 is a showcase with a P1 triage level of *InTriage*. More results, case studies, analysis and discussions can be seen in appendix.

6 Conclusion

This study presents *InTriage*, an AI-driven multilingual TT system for alleviating ineffectiveness in PEC triage. By automating call transcription, extracting information, and providing real-time triage decision support, *InTriage* improves accuracy, and enhances resource allocation in emergency settings. *InTriage* offers a scalable solution for modernizing TT, ensuring timely care for high-acuity cases and alleviating the burden on call-takers.

7 Ethics and Broader Impact Statement

This study was conducted in compliance with ethical guidelines to ensure the protection of participants' rights, privacy, and well-being. The research involved the collection, analysis, and use of anonymized data to improve triage protocols and enhance emergency response systems.

The data used in this study were fully anonymized before analysis to protect the privacy of individuals. No personally identifiable information was accessed or stored during the research process. Anonymization techniques ensured that participants could not be identified directly or indirectly from the data. Measures were taken to prevent potential biases in the data and to ensure that the AI system's recommendations adhered to the protocols established by the Singapore Civil Defence Force (SCDF). Regular monitoring and audits were implemented to assess the accuracy and fairness of the AI outputs. The study adhered to ethical standards regarding the use of AI technologies in healthcare. The AI system was designed to assist and augment human decision-making rather than replace it. Ethical safeguards were put in place to prevent harm and ensure the system's use aligns with SCDF's protocols.

Beyond ensuring privacy and data anonymization, we foresee two additional ethical risks that must be managed as InTriage scales: (1) over-reliance on automated recommendations and (2) inadvertent propagation of bias.

(1) Although our platform is designed to support, rather than replace human call-takers, the convenience brought by automation may gradually lead to over-reliance, potentially weakening call-takers' independent judgment and critical thinking. This risk is especially pronounced in emergency-care contexts when the system provide incorrect result. In our application scenarios, callers often speak incoherently under emotional distress and in noisy environments—conditions that may degrade ASR performance and further affect the system's overall reliability. Without appropriate safeguards, such over-reliance in high-stakes, unpredictable scenarios may compromise the accuracy and timeliness of medical responses, the very outcomes the system is intended to improve.

To curb over-reliance, we plan to clearly label all InTriage outputs as advisory only, reinforcing that the system is meant to support, not replace—human decision-making. Final clinical and dispatch decisions will remain the responsibility of trained call-takers. Furthermore, we aim to implement regular refresher training programs that help call-takers interpret model confidence scores, recognize edge cases where automation may be unreliable, and apply override rules when necessary. These training sessions will include scenario-based simulations that expose call-takers to challenging cases where the system might fail or produce uncertain outputs, thereby cultivating vigilance and situational awareness. Additionally, we intend to integrate real-time uncertainty indicators into the interface, such as low-confidence alerts, flagging ambiguous inputs, or highlighting cases with degraded ASR quality. These indicators will help users critically assess system recommendations rather than accept them at face value. Over time, we will also explore adaptive user interfaces that modulate the level of automation based on contextual reliability, for instance, reducing system assertiveness when acoustic conditions are poor or when user hesitation is detected. Ultimately, promoting a culture of active human oversight, where AI is seen as a collaborative tool rather than an infallible authority.

To counter potential biases from imbalanced training data, such as under-representation of non-standard accents or minority medical histories, we intend to conduct quarterly fairness audits across demographic subgroups and incident types. When disparities exceed predefined thresholds, we will apply targeted data augmentation or model fine-tuning. An internal ethics board comprising clinicians, technologists, and community representatives will review each audit and approve any algorithmic updates before deployment, ensuring continuous human oversight and accountability throughout the system's life cycle.

8 Acknowledgment

This project was funded by the National Research Foundation Singapore, Singapore under under AI Singapore Programme (Award Number: AISG2-TC-2022-004); by National University of Singapore, Saw Swee Hock School of Public Health, AI for Public Health (AI4PH) program, and The National Social Science Fund of China under Grant (23CWW006).

References

Christine Adler, Marion Krüsmann, Thomas Greiner-Mai, Anton Donner, Javier Mulero Chaves, and Àngels Via Estrem. 2011. It-supported management of mass casualty incidents: The e-triage project. In *Proceedings of the 8th International ISCRAM Conference*.

Mike Brady. 2020. Patient experiences of uk ambulance service telephone triage: a review of the literature. *International Journal of Emergency Services*, 9(2):89–108.

Min-Chen Chen, Ting-Yun Huang, Tzu-Ying Chen, Panchanit Boonyarat, and Yung-Chun Chang. 2023a. Clinical narrative-aware deep neural network for emergency department critical outcome prediction. *Journal of Biomedical Informatics*, 138:104284.

Yu-Hsuan Jamie Chen, Chin-Sheng Lin, Chin Lin, Dung-Jang Tsai, Wen-Hui Fang, Chia-Cheng Lee, Chih-Hung Wang, and Sy-Jou Chen. 2023b. An ai-enabled dynamic risk stratification for emergency department patients with ecg and cxr integration. *Journal of Medical Systems*, 47(1):81.

Hamza Elhaj, Nebil Achour, Marzia Hoque Tania, and Kurtulus Aciksari. 2023. A comparative study of supervised machine learning approaches to predict patient triage outcomes in hospital emergency departments. *Array*, 17:100281.

Ru Ying Fong, Wee Sern Sim Glen, Ahmad Khairil Mohamed Jamil, Wilson Wai San Tam, and Yanika Kowitlawakul. 2018. Comparison of the emergency severity index versus the patient acuity category scale in an emergency setting. *International Emergency Nursing*, 41:13–18.

Thomas Greiner-Mai and Anton Donner. 2010. Data management in mass casualty incidents: The e-triage project. In *INFORMATIK 2010 - 40. Jahrestagung der Gesellschaft für Informatik e.V.*, pages 200–206, Berlin, Germany.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2025. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Information Fusion*, 118:102963.

Kai He, Jialun Wu, Xiaoyong Ma, Chong Zhang, Ming Huang, Chen Li, and Lixia Yao. 2019. Extracting kinship from obituary to enhance electronic health records for genetic research. In *Proceedings of the Fourth social media mining for health applications (# SMM4H) workshop & shared task*, pages 1–10.

Thongpitak Huabbangyang, Rapeeporn Rojsaengroeng, Gawin Tiyawat, Agasak Silakoon, Alissara Vanichkulbodee, Jiraporn Sri-On, and Siriwimol Buathong. 2023. Associated factors of under and over-triage based on the emergency severity index; a retrospective cross-sectional study. *Archives of academic emergency medicine*, 11(1).

Ryota Inokuchi, Masao Iwagami, Yu Sun, Ayaka Sakamoto, and Nanako Tamiya. 2022. Machine learning models predicting undertriage in telephone triage. *Annals of Medicine*, 54(1):2989–2996.

Varada Vivek Khanna, Krishnaraj Chadaga, Niranjana Sampathila, Srikanth Prabhu, and Rajagopala Chadaga. 2023. A machine learning and explainable artificial intelligence triage-prediction system for covid-19. *Decision Analytics Journal*, 7:100246.

Qika Lin, Yifan Zhu, Xin Mei, Ling Huang, Jingying Ma, Kai He, Zhen Peng, Erik Cambria, and Mengling Feng. 2025. Has multimodal learning delivered universal intelligence in healthcare? a comprehensive survey. *Information Fusion*, 116:102795.

Jiafa Lu, Xin Wang, Linghao Chen, Xuedong Sun, Rui Li, Wanjing Zhong, Yajing Fu, Le Yang, Weixiang Liu, and Wei Han. 2023. Unmanned aerial vehicle based intelligent triage system in mass-casualty incidents using 5g and artificial intelligence. *World journal of emergency medicine*, 14(4):273.

Joshua R Lupton, Cynthia Davis-O'Reilly, Rebecca M Jungbauer, Craig D Newgard, Mary E Fallat, Joshua B Brown, N Clay Mann, Gregory J Jurkovich, Eileen Bulger, Mark L Gestring, et al. 2023. Undertriage and over-triage using the field triage guidelines for injured patients: a systematic review. *Prehospital Emergency Care*, 27(1):38–45.

Fateme Mohammadi, Ali Khani Jeihooni, Parisa Sabetsarvestani, Fozieh Abadi, and Mostafa Bijani. 2022. Exploring the challenges to telephone triage in prehospital emergency care: a qualitative content analysis. *BMC Health Services Research*, 22(1):1195.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Ma Jie, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, Stefano Soatto, et al. 2021. Structured prediction as translation between augmented natural languages. In *ICLR 2021-9th International Conference on Learning Representations*, pages 1–26. International Conference on Learning Representations, ICLR.

Ju Young Park. 2021. Real-time monitoring electronic triage tag system for improving survival rate in disaster-induced mass casualty incidents. In *Healthcare*, 7, page 877. MDPI.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

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

Singapore Civil Defence Force. 2023. Scdf annual statistics 2023. Accessed: 2024-12-19.

Andrew AS Soltan, Samaneh Kouchaki, Tingting Zhu, Dani Kiyasseh, Thomas Taylor, Zaamin B Hussain, Tim Peto, Andrew J Brent, David W Eyre, and David A Clifton. 2021. Rapid triage for covid-19 using routine clinical data for patients attending hospital: development and prospective validation of an artificial intelligence screening test. *The Lancet Digital Health*, 3(2):e78–e87.

Veronica Vicente, Maaret Castren, Fredrik Sjöstrand, and BIRGITTA WIREKLINT SUNDSTRÖM. 2013. Elderly patients' participation in emergency medical services when offered an alternative care pathway. *International Journal of Qualitative Studies on Health and well-being*, 8(1):20014.

Naythan Villafuerte, Santiago Manzano, Paulina Ayala, and Marcelo V García. 2023. Artificial intelligence in virtual telemedicine triage: A respiratory infection diagnosis tool with electronic measuring device. *Future Internet*, 15(7):227.

Jialun Wu, Xinyao Yu, Kai He, Zeyu Gao, and Tieliang Gong. 2024. Promise: A pre-trained knowledge-infused multimodal representation learning framework for medication recommendation. *Information Processing & Management*, 61(4):103758.

Yi Xiao, Jun Zhang, Cheng Chi, Yuqing Ma, and Aiguo Song. 2023. Criticality and clinical department prediction of ed patients using machine learning based on heterogeneous medical data. *Computers in Biology and Medicine*, 165:107390.

Guanrou Yang, Ziyang Ma, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024. Ctc-assisted llm-based contextual asr. *arXiv preprint arXiv:2411.06437*.

Zhuoyuan Yao, Di Wu, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei. 2021. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. In *Proc. Interspeech*, Brno, Czech Republic. IEEE.

A Appendix

A.1 Integration Challenges

Although InTriage convincingly demonstrates strong technical merit, navigating the complexities of deployment within a tightly regulated emergency care ecosystem remains a significant challenge. The most significant friction point is likely to come from government policy governing call monitoring and data sovereignty: transcribed audio and extracted medical data must comply with government policies. These rules impose strict limits on where audio can be stored, how long recordings may be retained, and which agencies may access them. These constraints could delay real-time hand-offs between InTriage, national EHR repositories, and ambulance-dispatch consoles. Also, call-takers may resist automation that appears to "shadow" or audit their work, especially if they fear increased scrutiny or deskilling. Overcoming these hurdles will require early engagement with regulators to codify permissible data flows, the introduction of opt-in call-monitoring policies that clarify accountability, and change-management programmes that position InTriage as a decision-support ally rather than a replacement. Without such policy alignment and workforce buy-in, even a well-validated AI pipeline risks limited uptake or protracted pilot phases within emergency services.

A.2 Training data

Table 3 provides a comprehensive summary of the training data used to develop the ASR agent of InTriage. The dataset is sourced from four primary audio types: Audiobooks, Podcasts, YouTube, and SCDF. Each source is characterized by varying acoustic conditions to ensure robustness across different environments and speaker profiles.

The Audiobook dataset comprises 2,655 transcribed hours and a total of 11,982 hours, capturing a wide range of reading voices across various ages and accents. Podcasts contribute 3,498 transcribed hours and a total of 9,254 hours, reflecting diverse conditions, including clean audio, background music, indoor and near-field recordings, and spontaneous speech. The YouTube dataset provides 3,845 transcribed hours and a total of 11,768 hours, covering both clean and noisy environments, with indoor, outdoor, near-field, and far-field recordings, as well as reading and spontaneous speech. Lastly, SCDF data includes 500 transcribed hours, comprising three years of real emergency call data from over 360,000 cases, providing real-world emergency scenarios. This diverse collection of audio data, with various acoustic conditions and speaker demographics, was essential in creating a robust ASR system capable of handling a wide range of real-world emergency call situations. At present, only transcribed data is used as train data. We will further enhance the ASR agent by employing more data in the future.

Audio Source	Transcribed Hours	Total Hours	Acoustic Condition
Audiobook	2,655	11,982	Reading, Various ages and accents
Podcast	3,498	9,254	Clean or background music, Indoor, Near-field, Spontaneous, Various ages and accents
YouTube	3,845	11,768	Clean and noisy, Indoor and outdoor, Near- and far-field, Reading and spontaneous, Various ages and accents
SCDF	500	-	3-years of real data from SCDF, more than 360,000 cases.

Table 3: Summary of used training data for ASR agent of *InTriage*.

Label	Precision (%)	Recall (%)	F1 (%)	Over-triage Rate (%)	Under-triage Rate (%)
P1	18.82 / 28.27	65.31 / 77.14	29.22 / 41.38	0.00 / 0.00	34.69 / 22.86
P2	80.81 / 85.90	54.79 / 59.45	65.31 / 70.27	44.52 / 39.94	0.68 / 0.61
P3	50.00 / 83.33	6.45 / 31.25	11.43 / 45.45	93.55 / 68.75	0.00 / 0.00
Macro-Average	49.88 / 65.84	33.52 / 55.63	32.77 / 52.03	–	–
Micro-Average	52.15 / 60.23	45.01 / 60.09	45.32 / 60.16	–	–
Avg.				46.02 / 36.23	11.79 / 7.82

Table 4: Triage performance breakdown. Black text indicates human performance, while red text represents the performance of our InTriage system.

A.3 Evaluation data

Table 6 presents the various s used to evaluate the performance of the ASR system across different language varieties, primarily focusing on Singaporean-English, Mandarin-English, and Malay-English. The s were carefully selected to cover a range of real-world scenarios to ensure robustness in different emergency communication contexts.

For Singaporean-English, multiple s were employed, including data from the Ministry of Social and Family Development, Singapore (MSF), MSF-DHL Express, Singapore (DHL), and NTU Inhouse recordings. Additionally, s from the Singapore Civil Defence Force (SCDF) were used, comprising two distinct s (SCDF testset (1) and SCDF testset (2)), reflecting emergency call scenarios. The Infocomm Media Development Authority, Singapore (IMDA) testsets (1 and 2) were also included to assess performance in broader communication contexts. Notably, IMDA testset (2) Hotword Boosting was used to evaluate the ASR system's ability to recognize critical keywords and phrases in emergency situations.

For Mandarin-English and Malay-English, the s were sourced from SCDF, labeled as SCDF-Mandarin and SCDF-Malay respectively. These two datasets support language-mixing between Mandarin, Malay, and English languages. Under such conditions, it ensures that the ASR system is evaluated across a multilingual context with different accent, reflecting the linguistic diversity of Singapore's population and the need for multilingual support in emergency response systems. The comprehensive inclusion of these s highlights the ASR system's capacity to handle various accented English varieties and language-switched speech, ensuring reliable performance in multilingual, spontaneous, and formal communication settings.

Table 4 provides a detailed performance comparison between human call-takers and our proposed InTriage system across various triage categories. Overall, InTriage demonstrates substantial improvements in critical performance metrics. For instance, in category P1, our system notably increases precision from 18.82% to 28.27% and recall from 65.31% to 77.14%, resulting in an improvement in the F1 score from 29.22% to 41.38%. Even more pronounced improvements are observed in category P3, with precision significantly rising from 50.00% to 83.33%, recall increasing from 6.45% to 31.25%, and consequently, the F1 score markedly improving from 11.43% to 45.45%.

The results further indicate that while both human call-takers and InTriage exhibit relatively low under-triage rates, over-triage remains a notable challenge. This primarily stems from the conservative, safety-first approach generally adopted by human call-takers, as evidenced by the particularly high over-triage rates observed in categories P2 and P3. Even in cases not clearly urgent, human call-

<table>
<tr><th colspan="6">Case 1</th></tr>
<tr><td>ASR
(part results)</td><td colspan="5">...
Operator: oh did she happen to choke on anything
Caller: oh, no i am not sure
Operator: uh if she speaks to you is it in full sentence or word by word
Caller: half half yup correct
Operator: can I say word by word
Caller: ya just word by word
...</td></tr>
<tr><td rowspan="2">NLU</td><td>Name</td><td>Gender</td><td>Age</td><td>Location</td><td>Incidence</td></tr>
<tr><td>[mask]</td><td>female</td><td>66</td><td>[mask]</td><td>breathing
problem</td></tr>
<tr><td>Triage</td><td>P1+</td><td>SA</td><td>neural</td><td colspan="2"></td></tr>
<tr><td>Evidence</td><td colspan="5">breathing problem → when s/he speaks, is it in full sentence or only one word at a time ? : one/few words</td></tr>
<tr><td rowspan="5">Details</td><td colspan="5">is s/he choking now? : no (unknown)</td></tr>
<tr><td colspan="5">can s/he respond in the usual way when you call /tap (alert)? : not given</td></tr>
<tr><td colspan="5">when s/he speaks, is it in full sentence or only one word at a time (long/short cry)? : One/few words</td></tr>
<tr><td colspan="5">is s/he coughing blood? : not given</td></tr>
<tr><td colspan="5">does s/he have asthma or any lung disease? : not given</td></tr>
<tr><th colspan="6">Case 2</th></tr>
<tr><td>ASR
(part results)</td><td colspan="5">...
Operator: is there any blood
Caller: no (oh), on my head, i feel very painful
Operator: if anyone trapped in the car
Caller: no no one trapped in the car
...</td></tr>
<tr><td rowspan="2">NLU</td><td>Name</td><td>Gender</td><td>Age</td><td>Location</td><td>Incidence</td></tr>
<tr><td>[mask]</td><td>-</td><td>-</td><td>[mask]</td><td>motor
vehicle
accident</td></tr>
<tr><td>Triage</td><td>P3 (P2)</td><td>SA</td><td>anxious</td><td colspan="2"></td></tr>
<tr><td>Evidence</td><td colspan="5">motor vehicle accident-> is there any blood? : no</td></tr>
<tr><td rowspan="6">Details</td><td colspan="5">what vehicle(s)is/are involved? : vehicle</td></tr>
<tr><td colspan="5">is anyone trapped (pinned) in/under the vehicle ? : no</td></tr>
<tr><td colspan="5">was anyone thrown (flung) from the vehicle?: not given</td></tr>
<tr><td colspan="5">is everyone involved in the accident able to walk (alert)?: not given</td></tr>
<tr><td colspan="5">are there any obvious injuries? : not given</td></tr>
<tr><td colspan="5">is there any blood? : no (yes)</td></tr>
</table>

Table 5: Case study of two flawed examples. Erroneous elements are highlighted in red, and corrections are provided within brown parentheses.

takers tend to dispatch ambulances as a precaution, leading to an extremely high over-triage rate of up to 93.55% in category P3. In contrast, InTriage, trained using hospital-derived goal triage labels, is less influenced by subjective biases. We anticipate that InTriage will provide additional objective guidance to assist human call-takers in reducing over-triage rates, thereby optimizing ambulance resource utilization.

A.4 Cases studies

Table 5 presents two representative error cases that illustrate the different ways in which component-level failures occur.

In Case 1, the ASR module accurately transcribes all salient utterances, allowing the evidence extractor to trigger on the critical "breathing prob-

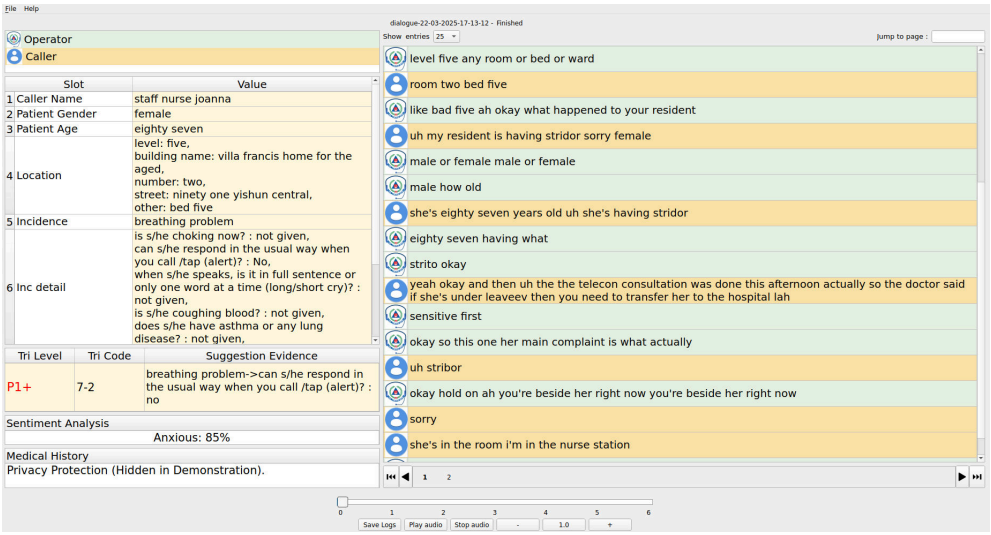

Figure 8: Screenshot of *InTriage* for an emergency case with P1+ level.

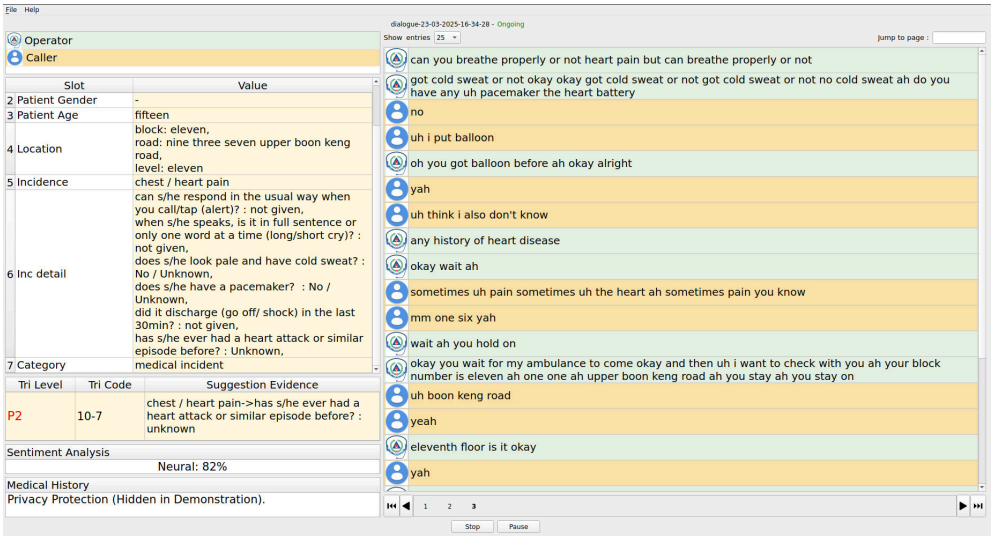

Figure 9: Screenshot of *InTriage* for an emergency case with P2 level.

lem" pattern. However, the NLU agent misidentifies "no i am not sure" as "no" (highlighted in red; correct value shown in brown parentheses). Because the caller's description is somewhat ambiguous and includes the literal expression "no," the error is understandable. Nevertheless, it did not affect the accuracy of the final triage outcome. This example highlights the system's robustness to NLU errors in non-critical fields when key clinical triggers are correctly identified.

In Case 2, the caller confirms the presence of bleeding after a motor-vehicle collision, but ASR mistranscribes the response as "no" instead of "oh". This may be due to the similar pronunciation of the two words in a noisy background. Consequently, NLU extracts the feature no bleeding, and the inference module downgrades the incident from the correct P2 (potentially life-threatening) to P3 (less urgent). This error highlights how a single misrecognized word can sometimes lead to an incorrect final triage outcome. As such, there is still significant room for further research in this area. This also underscores why we position our system as decision support rather than decision making.

A.5 More show cases

Figure 8, 9, 10 provide three examples with correct results for P2, P3, and P4 triage levels. Based on the different triage levels, the call-taker will make the final decision on whether to dispatch an ambulance and determine the appropriate type of ambulance to be dispatched.

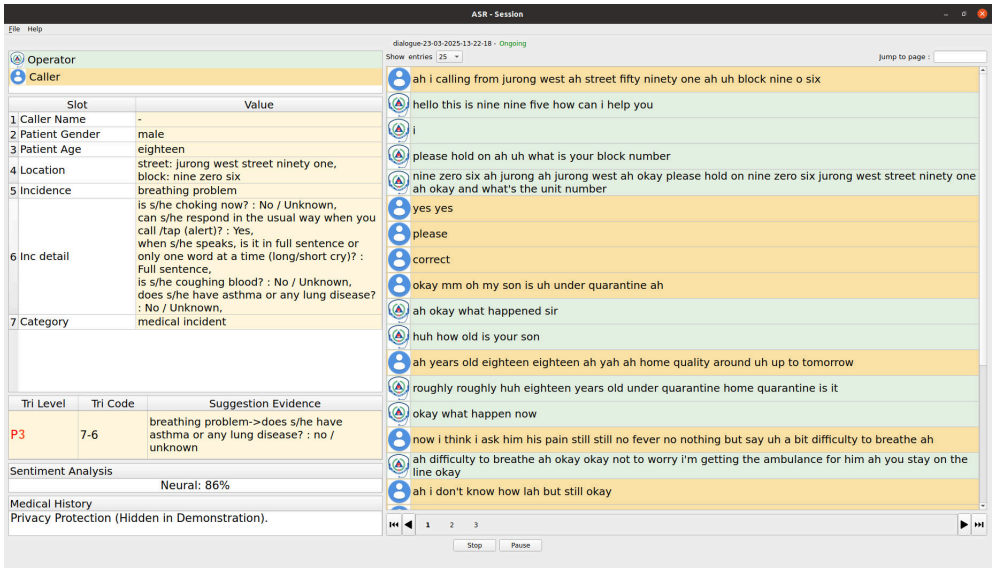

Figure 10: Screenshot of *InTriage* for an emergency case with P3 level.