Multilingual, Dialect-Robust Language Technologies for Safe Sequential Triage

Dipankar Srirag

1 Introduction

Triage, the first step of a patient's visit to a hospital emergency department (ED), centres on a conversation. The triage nurse elicits the chief complaint, the patient's main reason for coming, gathers vital signs and history, and assigns an urgency level from 1 (most urgent) to 5 (least urgent) on the Australasian Triage Scale (ATS; Australasian College for Emergency Medicine, 2023) or the US Emergency Severity Index (ESI; Gilboy et al., 2005). Triage is *sequential*, because evidence arrives turn by turn, and *time-constrained*, because the more critical the patient, the less time the nurse has to decide. Yet most evaluations of large language models (LLMs) for triage judge them on a written-up case, the *outcome* of an interaction, rather than on the evidence available *during* it (**Srirag et al.**, 2026a). To evaluate triage as it unfolds, I define *sequential clinical triage* as predicting the urgency level from a growing prefix of the nurse–patient conversation.

A system deployed in an ED must serve multilingual patients, for whom language barriers are linked to underestimated urgency (Rojas et al., 2023), and fit clinical workflows, including hospital infrastructure, privacy rules (**Srirag et al.**, 2026a) and the nurse's final decision. Such a system is planned to be tested in three Sydney EDs with patients who speak Arabic, the most spoken language in South Western Sydney after English (Cancer Institute NSW, 2023). **This motivates my PhD (Figure 1), which develops a language technology for sequential triage that preserves evidence across languages and dialects, updates its prediction as evidence arrives and is designed for safe use in an ED.**

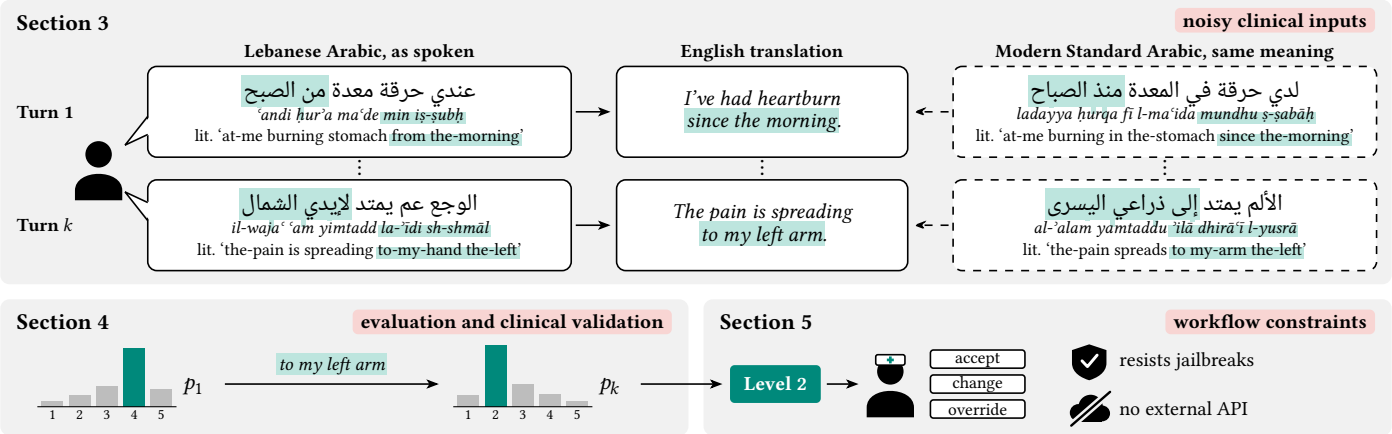


Figure 1: The running example at turn 1 and a later turn k, where p_t is the system's distribution over the five urgency levels after turn t. The patient's Lebanese Arabic differs from Modern Standard Arabic word for word but has the same English translation. Highlighted text marks the evidence each turn adds. Red boxes name the open challenge each section addresses.

2 Prior Work

My prior work covers dialect robustness, translation across languages and the evaluation of sequential triage.

The accuracy of language models in understanding conversations drops in a dialect (**Srirag et al.**, 2025a), and a *low-rank dialect adapter*, trained with a contrastive loss between standard and dialectal representations, improves it (**Srirag et al.**, 2025b). BESSTIE (**Srirag et al.**, 2025c), my sentiment and sarcasm benchmark for three English dialects, shows that models fine-tuned on one dialect lose accuracy on another, so each dialect needs its own data. DiaLLM (Painter et al., 2026) shows a gap between understanding a dialect and producing it, which matters for a system that replies to patients in their dialect. In a white paper for the NSW Department of Education, I evaluated LLM translation of school communication across 22 languages.[1] I analysed its errors and its trade-off between cost and accuracy. This motivates my work on dialect-robust translation of triage conversations (Section 3).

My survey (**Srirag et al.**, 2026a) organises NLP in the ED by the three phases of emergency care, from triage to discharge, and I teach it as a tutorial at AJCAI 2026.[2] This thesis addresses three open challenges that the survey identifies, **noisy clinical inputs**, **evaluation and clinical validation** and **workflow constraints**. Because regulation

[1] https://apo.org.au/node/331605

[2] https://ajcai26.org/workshopandtutorial.html

limits triage research to structured health records, I built TriageSim (**Srirag et al.**, 2026b), a *simulation framework* that turns these records into nurse–patient conversations with audio in four accents; a public release is under review on PhysioNet. Using simulated and clinician-authored conversations, I showed that every LLM tested *anchors on the chief complaint* when it triages turn by turn and falls short of clinicians (**Srirag et al.**, 2026c). Evaluation on completed records hides this failure, and prompting does not fix it. This motivates my planned work on measuring and reducing anchoring (Section 4).

3 Dialect-Robust Translation of Triage Conversations

Imagine a patient who speaks Lebanese Arabic and tells the nurse that she has had heartburn since the morning.

Spoken Arabic dialects differ from the standard written language in vocabulary and grammar (Zaidan and Callison-Burch, 2014), and such language barriers are one source of **noisy clinical inputs**. Omissions are the most common error made by medical interpreters (Flores et al., 2003). An omitted detail, such as when her pain began, never reaches the triage model. Translation models can learn from quality feedback rather than from reference translations (He et al., 2024), which are scarce for dialectal Arabic (El Mekki et al., 2026). In ongoing work, I develop LLM-based translation between English and dialectal Arabic that learns from feedback on dialect, clinical content and implied meaning. To measure whether translation preserves a conversation's urgency level, I propose to extend TriageSim to dialectal English and dialectal Arabic. The same conversations allow a comparison of triage after translation with triage in the source language. Since triage is time-constrained, a further question is how simultaneous translation, which translates while the patient speaks (Ma et al., 2019), trades latency against accuracy.

4 Measuring and Reducing Anchoring in Sequential Triage

Several turns later, the same patient mentions that the pain spreads to her left arm.

This late warning sign should raise her urgency, yet LLMs pick up such details without using them to revise their estimate (**Srirag et al.**, 2026c). A similar failure appears outside medicine, where LLMs make assumptions in early turns and do not recover (Laban et al., 2025). Writing p_t for a model's distribution over the five urgency levels after turn t, I propose to measure how far p_t should move from p_{t-1} as the information turn t adds about the correct level, and to compare models' shifts with clinicians', which addresses **evaluation and clinical validation**. Methods that track an explicit patient state that each turn revises (Mrkšić et al., 2017) or weigh each turn by the information it adds may reduce anchoring. I will test them on TriageSim conversations in which a late turn changes the correct level.

5 Compact, Jailbreak-Resistant Models with Nurses in the Loop

The system then suggests a level for this patient, and the nurse decides.

Published work reports no language technology deployed in an ED workflow (**Srirag et al.**, 2026a), and a model's accuracy in isolation does not carry over to the decisions of the people who use it (Bean et al., 2026). Among **workflow constraints**, sending her conversation to the system's current models, which are served through external APIs, conflicts with health privacy requirements (**Srirag et al.**, 2026a). In its replies to her, the system must neither give harmful advice nor reveal other patients' details. I propose to explore compact models, built with low-rank adaptation and quantisation, that need no external API. In upcoming work, I will refine an open foundation model for the medical domain to resist jailbreaks that elicit harmful advice or private details, while preserving its clinical usefulness. Since non-English prompts can bypass the safety training of LLMs (Deng et al., 2024), the compact models must be tested against the same attacks phrased in dialectal Arabic or dialectal English. Finally, a usability study in Sydney EDs will compare the system with standard triage with interpreter services, which extends the **evaluation and clinical validation** of Section 4. The study will examine how nurses accept, change or override the system's suggestions, since time pressure invites over-reliance on automated suggestions (Goddard et al., 2012).

6 Applications Beyond Triage

Dialect-robust translation can serve legal aid and customer support for speakers of dialects. Simultaneous translation trades latency against accuracy, as do systems that act before all the evidence has arrived, such as fraud detection and trading. Measuring whether a model's updates match the information each turn adds applies to LLM agents that work with users over many turns (Laban et al., 2025). Compact models that resist jailbreaks in dialects suit domains where data cannot leave the organisation, such as law and finance. I aim to build language technology that works across dialects and updates its predictions as evidence arrives.

References

Australasian College for Emergency Medicine. 2023. Policy on the Australasian Triage Scale (P06, version 5).

Andrew M. Bean, Rebecca Elizabeth Payne, Guy Parsons, Hannah Rose Kirk, Juan Ciro, Rafael Mosquera-Gómez, Sara Hincapié M, Aruna S. Ekanayaka, Lionel Tarassenko, Luc Rocher, and Adam Mahdi. 2026. Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. *Nature Medicine*, 32(2):609–615.

Cancer Institute NSW. 2023. Multicultural Demographics Data Explorer. NSW Local Health Districts: Top 10 Birthplace and Language Groups Other Than English. Data from the 2021 Census of Population and Housing.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. Multilingual Jailbreak Challenges in Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Abdellah El Mekki, Samar M. Magdy, Houdaifa Atou, et al. 2026. Alexandria: A Multi-Domain Dialectal Arabic Machine Translation Dataset for Culturally Inclusive and Linguistically Diverse LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32567–32592, San Diego, California, United States. Association for Computational Linguistics.

Glenn Flores, M. Barton Laws, Sandra J. Mayo, Barry Zuckerman, Milagros Abreu, Leonardo Medina, and Eric J. Hardt. 2003. Errors in Medical Interpretation and Their Potential Clinical Consequences in Pediatric Encounters. *Pediatrics*, 111(1):6–14.

Nicki Gilboy, Paula Tanabe, Debbie A. Travers, Alexander M. Rosenau, and David R. Eitel. 2005. *Emergency Severity Index, Version 4: Implementation Handbook*. Agency for Healthcare Research and Quality, Rockville, MD. AHRQ Publication No. 05-0046-2.

Kate Goddard, Abdul Roudsari, and Jeremy C. Wyatt. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127.

Zhiwei He, Xing Wang, Wenxiang Jiao, Zhuosheng Zhang, Rui Wang, Shuming Shi, and Zhaopeng Tu. 2024. Improving Machine Translation with Human Feedback: An Exploration of Quality Estimation as a Reward Model. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8164–8180, Mexico City, Mexico. Association for Computational Linguistics.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. LLMs Get Lost In Multi-Turn Conversation. *arXiv preprint arXiv:2505.06120*.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous Translation with Implicit Anticipation and Controllable Latency using Prefix-to-Prefix Framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural Belief Tracker: Data-Driven Dialogue State Tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Vancouver, Canada. Association for Computational Linguistics.

Jordan Painter, **Dipankar Srirag**, Adarsh Kappiyath, Diptesh Kanojia, Aditya Joshi, and Lu Yin. 2026. DiaLLM: An Investigation into the Robustness-Generation Gap in English Dialect Adaptation. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, pages 52728–52745, Budapest, Hungary. Association for Computational Linguistics.

Christina R. Rojas, James M. Chamberlain, Joanna S. Cohen, Gia M. Badolato, Lawrence J. Cook, Monika K. Goyal, and Deena Berkowitz. 2023. Undertriage for Children With Caregivers Preferring Languages Other Than English. *Pediatrics*, 151(6):e2022059386.

Dipankar Srirag, Aditya Joshi, and Jacob Eisenstein. 2025b. Predicting the Target Word of Game-playing Conversations using a Low-Rank Dialect Adapter for Decoder Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 8–17, Albuquerque, New Mexico. Association for Computational Linguistics.

Dipankar Srirag, Aditya Joshi, Salil S. Kanhere, and Padmanesan Narasimhan. 2026a. From Triage to Discharge: A Survey of NLP Tasks, Methods, and Open Challenges in the Emergency Department. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, pages 20918–20942, Budapest, Hungary. Association for Computational Linguistics.

Dipankar Srirag, Aditya Joshi, Jordan Painter, and Diptesh Kanojia. 2025c. BESSTIE: A Benchmark for Sentiment and Sarcasm Classification for Varieties of English. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8413–8429, Vienna, Austria. Association for Computational Linguistics.

Dipankar Srirag, Quoc Dung Nguyen, Aditya Joshi, Padmanesan Narasimhan, and Salil Kanhere. 2026b. TriageSim: A Conversational Emergency Triage Simulation Framework from Structured Electronic Health Records. In *Interspeech 2026*, pages 1997–2001.

Dipankar Srirag, Nihar Ranjan Sahoo, and Aditya Joshi. 2025a. Evaluating Dialect Robustness of Language Models via Conversation Understanding. In *Proceedings of the Second Workshop on Scaling Up Multilingual & Multi-Cultural Evaluation*, pages 24–38, Abu Dhabi. Association for Computational Linguistics.

Dipankar Srirag, Haokai Zhao, Ashutosh Kumar, Eleanor Hopper, Michael Dalton, Quoc Dung Nguyen, Aditya Joshi, Salil S. Kanhere, and Padmanesan Narasimhan. 2026c. LLMs Anchor on Chief Complaint and Fail to Integrate Evidence in Sequential Clinical Triage. *arXiv preprint arXiv:2609.22904*.

Omar F. Zaidan and Chris Callison-Burch. 2014. Arabic Dialect Identification. *Computational Linguistics*, 40(1):171–202.