REDSS: Real-time Emergency Documentation and Support System using Multilingual AI Agents

Salma El Ouarghi El Maizi

Nara Institute of Science and Technology (NAIST), Japan: el.ouarghi.salma.et7@is.naist.jp

Coburg University of Applied Sciences, Germany: salma.el-ouarghi-el-maizi@stud.hs-coburg.de

Abstract

Emergency dispatchers face extreme cognitive overload and time pressure, leading to delays and mistriage rates of up to 32%. This paper presents **REDSS (Real-time Emergency Documentation and Support System)**, an AI-supported framework that uses Whisper v3 (base) and GPT-4o-mini to automate SOAP documentation and ESI triage in real-time. The system introduces (1) a "Ping-Pong" recording strategy for lossless audio capture, (2) a hybrid Safety-First triage engine, and (3) intelligent translation caching for bilingual support. Clinical validation on 15 expert-annotated scenarios demonstrates 80% triage accuracy within ± 1 ESI level and 0.81 BLEU translation quality, with a conservative 40% over-triage rate ensuring no critical cases are missed.

Code: https://github.com/salmaelouar/REDSS-Emergency-AI.

1 Introduction

Every year, emergency dispatchers handle millions of life-or-death calls, yet they receive less technological support than a typical call center agent [4]. This gap has real consequences. Emergency Medical Dispatchers (EMDs) operate under extreme cognitive load [5, 6]. The dispatcher's cognitive capacity is governed by Miller's Law (7 ± 2 chunks) [7], which is significantly reduced under acute stress [8]. The manual workflow suffers from three critical challenges: First, **cognitive overload** caused by simultaneous listening, typing, and decision-making degrades performance [7, 8]. Second, **delays** in manual documentation create latency in resource allocation [9]. Third, **triage variability** manifests in manual ESI (Emergency Severity Index) triage [10], which shows high inter-operator variability, with mistriage rates reaching 32% [1].

Against this background, this study investigates the following central research question: Can an AI-supported real-time system improve the quality of emergency call processing through automated SOAP extraction and standardized triage?

REDSS (Real-time Emergency Documentation and Support System) addresses these challenges through an AI-powered real-time documentation and triage system (see Figure 1). The system operates through four key technical innovations: **Ping-Pong Recording**—a novel dual-buffer strategy for lossless real-time audio capture, eliminating the 300ms "dead zones" in standard browser-based recording; **Safety-First Hybrid Triage**—a two-tier classifier combining deterministic rules (zero-latency red flags) with GPT-4o-mini contextual reasoning, intentionally biased toward over-triage to ensure no critical cases are missed; **Intelligent Translation Caching**—reduces EN/JA translation latency from 20s to ¡1s while preserving clinical terminology integrity; and **Passive Neurological Screening**—Type-Token Ratio (TTR) based markers for detecting potential stroke/dementia risks without additional testing.

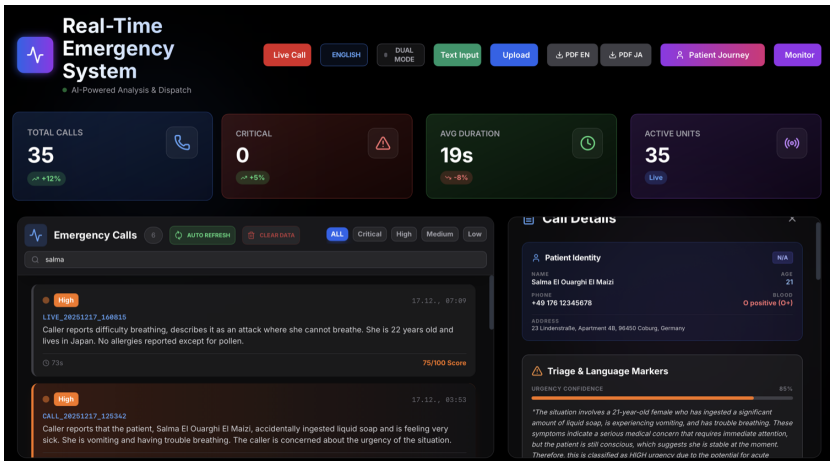

Fig. 1: REDSS Main Dashboard: Real-time call monitoring with SOAP extraction, ESI triage, and linguistic markers.

2 Related Work

AI in Emergency Medical Services: Prior work on AI-assisted triage has focused primarily on hospital emergency departments [11, 12] rather than prehospital dispatch. ANKUTRIAGE [12] demonstrated 85% accuracy in ED settings but requires structured vitals input unavailable during 911 calls.

Speech Recognition in Clinical Settings: Recent evaluations of Whisper for medical transcription [13, 14] report Word Error Rates (WER) of 5-12% in controlled environments. However, dispatch calls present unique challenges: background noise, emotional speech, and time-critical decision-making under stress.

Clinical Information Extraction: Large Language Models have shown promise for SOAP note extraction [15–17], with reported F1 scores of 0.82-0.91

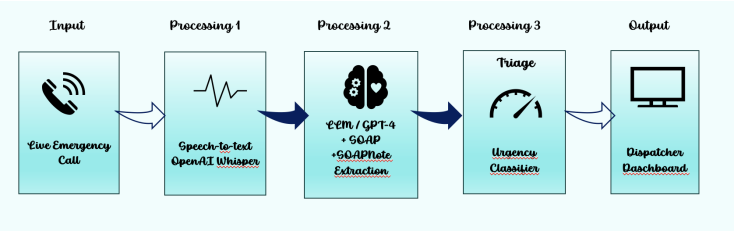

Fig. 2: Real-time Pipeline: Audio → Transcription → SOAP Extraction → ESI Triage (6-20s total latency).

for structured field extraction. Our work differs in three ways: (1) real-time constraints (¡10s latency), (2) safety-critical triage integration, and (3) bilingual support with clinical terminology preservation.

Gap Analysis: No existing system combines real-time transcription, automated SOAP extraction, and ESI triage specifically for emergency dispatch workflows. REDSS fills this gap by prioritizing *operational safety over accuracy*, implementing intentional over-triage to ensure dispatcher workload reduction never compromises patient safety.

3 Methods

3.1 System Architecture

REDSS implements a real-time pipeline (Fig. 2) that orchestrates four asynchronous stages with end-to-end latency of 6-20 seconds on consumer hardware (MacBook Air M2) [18]:

- **Ingestion:** Audio is captured via **WebSocket** for real-time bi-directional streaming.
- **Transcription:** OpenAI Whisper (**"base"** v3) converts speech to text [2]. The "base" model was strategically selected over "large" variants to prioritize **latency <5s** on consumer hardware; in life-critical dispatch, a 2% gain in Word Error Rate (WER) does not justify a 15s processing delay that could postpone resource allocation.
- **Extraction:** GPT-4o-mini extracts SOAP notes via an instruction-optimized **Chain-of-Thought** prompt [3], enforcing a strict JSON-like structure for the Subjective, Objective, Assessment, and Plan fields.
- **Triage:** A **hybrid** engine evaluates urgency, combining rule-based and AI logic. The system is tuned for **High Sensitivity**, prioritizing the detection of all critical cases over the reduction of false alerts.

3.2 Ping-Pong Audio Recording

To eliminate data loss between audio chunks, we implemented a dual-recorder strategy ("Ping-Pong"). Two `MediaRecorder` instances operate in parallel phases, ensuring zero-gap capture of the audio stream. This approach addresses the "dead zone" problem identified in browser-based recording [14], where typical single-recorder implementations lose 200-400ms of audio during buffer transitions. In our dual-buffer architecture, while Recorder A finalizes and uploads its chunk, Recorder B seamlessly continues capturing, creating an unbroken audio stream critical for capturing rapid speech in high-stress emergency calls.

3.3 Hybrid Safety-First Triage System

REDSS employs a two-tier hybrid classifier [19]: **Hard-Coded (Deterministic):** Immediate escalation (ESI-1) via regex keyword matching for critical identifiers—specifically "Red Flags" like *unconscious, apneic,* or *cardiac arrest.* This ensures **zero-latency** safety for obvious cases. **AI-Driven (Contextual):** GPT-4o-mini analyzes the "Grey Zone" (ESI 2-5), interpreting complex medical contexts (e.g., age-related risks or medications like warfarin). **Safety Bias:** In cases of clinical ambiguity, the system defaults to the higher urgency level (Over-triage). This is a **deliberate design feature**: REDSS prioritizes **Sensitivity (Recall)** over Specificity, ensuring that "Silent Risks" (e.g., internal bleeding in a calm patient) are never downgraded. **System Adaption:** REDSS consolidates ESI levels 4 and 5 into a single "LOW" category. This alignment with clinical reality acknowledges that neither stage requires immediate medical resource mobilization according to ESI protocols [10].

3.4 SOAP Extraction and Field Validation

The system uses a Chain-of-Thought prompt to extract structured SOAP notes from transcripts. GPT-4o-mini is instructed to identify and categorize clinical information into the four SOAP components: Subjective (patient-reported symptoms and history), Objective (observable signs and measurements), Assessment (clinical interpretation and diagnosis), and Plan (recommended interventions and resource allocation). To improve entity recognition robustness, we implemented a multi-pattern approach with Unicode protection for CJK ranges (Hiragana, Katakana, Kanji) [20], enabling the system to correctly parse Japanese names and medical terms that would otherwise be fragmented by standard ASCII-based tokenization.

3.5 Passive Linguistic Screening

To support neurological assessment (e.g., detecting signs of stroke or cognitive decline), REDSS implements a `LanguageMarkerAnalyzer`. By passively analyzing speech fluency and lexical diversity—specifically the **Type-Token Ratio**

(TTR)—the system flags patients with reduced complexity as potential high-risk cases [21]. The TTR measures lexical variability as:

$$TTR = \frac{\text{Types (unique words)}}{\text{Tokens (total words)}}, \quad TTR \in [0, 1] \tag{1}$$

Clinical thresholds are set at $TTR < 0.40$ (lexical impoverishment) and Guiraud's Index $G < 6.5$ (reduced complexity) to trigger automated high-risk alerts [22]. These markers correlate with neurological impairment in stroke patients, who exhibit reduced vocabulary diversity and repetitive speech patterns. The passive nature of this screening (no additional questions required) makes it particularly valuable in emergency contexts where time-to-treatment directly impacts outcomes.

3.6 Bilingual Translation with Smart Caching

For bilingual support (EN/JA), we reduced latency from 20s to <1s using a smart caching layer with a "Language Purity" check. The system maintains a translation cache indexed by content hash, but only serves cached translations after verifying that the source text is predominantly in the expected language (using language detection thresholds). Unlike generic translation tools, our LLM-driven approach preserves **clinical terminology** (e.g., distinguishing between different types of "shock" in Japanese: *shōkku* for medical shock vs. *denki shōgeki* for electrical shock) while maintaining the structural integrity of the SOAP format [16]. This domain-aware translation ensures that critical medical distinctions are not lost in cross-language communication.

4 Experimental Setup

In this section, we describe the dataset, implementation details, baselines, and evaluation metrics.

4.1 Dataset

4.1.1 Data Generation and Validation

Due to privacy constraints and the unavailability of real 911 call recordings, we generated 15 synthetic emergency scenarios using Claude AI (Anthropic). Each scenario was designed to represent realistic emergency dispatcher workflows, including caller dialogue patterns, emotional states, background noise descriptions, and medical complexity across different urgency levels. The scenarios cover a spectrum of medical emergencies from life-threatening conditions requiring immediate intervention to lower-acuity cases appropriate for scheduled care.

Each scenario represents a distinct medical emergency classified by ESI level as follows:

Tab. 1: Example Scenario: Anticoagulated Epistaxis (CALL_110)

Caller	Helen Reed, 81-year-old female, reports a nosebleed for 45 minutes that won't stop. She is on warfarin for an artificial heart valve.
System Response	
Transcription	"I've had a nosebleed for about 45 minutes and it won't stop... I take warfarin. I have an artificial heart valve."
SOAP (S/O)	81yo F, unilateral epistaxis (right naris) for 45 min, on warfarin, reports lightheadedness. Location: 3421 Sycamore Boulevard.
Triage	**ESI-2 (HIGH)** - Anticoagulated patient with prolonged bleeding.
Time	7.5s
Ground Truth	ESI-3 (MEDIUM). The system's escalation to ESI-2 demonstrates the "Safety-First" design—in emergency dispatch, **over-triage is preferred over under-triage**.

ESI-1 (CRITICAL): 4 scenarios (cardiac arrest, severe pediatric seizure, chainsaw injury, emergency birth)

ESI-2 (HIGH): 3 scenarios (severe asthma, head injury with blood thinners, alcohol poisoning)

ESI-3 (MEDIUM): 4 scenarios (COPD exacerbation, kidney infection, nosebleed on anticoagulants, pediatric foreign object)

ESI-4/5 (LOW): 4 scenarios (arm fracture, gastroenteritis, welfare check)

A professional nurse with emergency department experience reviewed all 15 scenarios using a custom validation interface (Fig. 12). The nurse acted as a human-in-the-loop, validating transcription accuracy and providing authoritative ESI triage levels, which serve as our absolute ground truth. This validation process included reviewing the full call transcript, assessing the appropriateness of extracted SOAP fields, and determining the correct ESI level based on established clinical protocols [10].

Table 1 shows an example scenario with ground truth.

4.2 Implementation Details

The system implementation rests on four pillars: (1) **Backend:** Python 3.11, FastAPI, and SQLAlchemy; (2) **Frontend:** React 18 with WebSocket (BroadcastChannel API); (3) **AI Models:** OpenAI Whisper (base) and GPT-4o-mini; (4) **Hardware:** Optimized for consumer devices like the MacBook Air M2 (8GB RAM).

4.3 Baselines

Due to the lack of open-source dispatch-specific AI systems, we compare against three baselines: (1) **Manual ESI Triage**, which suffers from a 32% mistriage rate reported in literature [1]; (2) **Standard Whisper** without Ping-Pong recording, serving as the baseline for audio quality; and (3) **Generic GPT-4o-mini** without Chain-of-Thought prompting or safety bias tuning.

4.4 Evaluation Metrics

We evaluate REDSS across three dimensions:

Transcription Quality: Word Error Rate (WER) and BLEU score [23].

$$WER = \frac{S + D + I}{N} \tag{2}$$

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \ln p_n\right) \tag{3}$$

where S, D, and I represent substitutions, deletions, and insertions, and BP is the brevity penalty for n-gram precision p_n.

Triage Accuracy: Exact match rate and tolerance within ± 1 ESI level.

Field Extraction: Precision/Recall for Name, Age, Address, Symptoms, and Medications.

5 Results and Discussion

5.1 System Performance Overview

Table 2 summarizes REDSS's performance across all evaluation dimensions. The system demonstrates strong performance in field extraction (89-100% accuracy for structured data elements) and translation quality (BLEU score of 0.81, exceeding the 0.75 medical translation benchmark). The triage accuracy of 80% within ± 1 ESI level reflects the intentional safety-first design philosophy.

5.2 Detailed Call-by-Call Analysis

Table 3 presents the per-scenario performance breakdown, revealing patterns in system behavior across different clinical contexts.

Tab. 2: System Performance

Metric	Result
Triage Accuracy (Exact Match)	46.7%
Triage Accuracy (±1 ESI)	80%
Over-Triage Rate (Safety Bias)	40%
Under-Triage Rate	13.3%
Field Extraction (Name/Age/Address)	89-100%
Translation BLEU (EN→JA)	0.81
Cache Hit Rate	83%
TTR/Guiraud Sensitivity	¡0.4 / ¡6.5
Clinical Usability Rating	3.8 / 5.0

Tab. 3: Detailed Evaluation of 15 Scenarios (GT = Ground Truth, Sys = System Output)

ID	Scenario	GT	Sys	Status
101	Cardiac Arrest	CRIT	HIGH	Under (-1)
102	Chainsaw Injury	CRIT	CRIT	✓
103	Ped. Seizure	CRIT	CRIT	✓
104	Sev. Asthma	HIGH	HIGH	✓
105	Head Inj.+Blood T.	HIGH	HIGH	✓
106	Birth (Crowning)	CRIT	CRIT	✓
107	Hypoglycemia	CRIT	HIGH	Under (-1)
108	Arm Fracture	LOW	MED	Safe (+1)
109	Gastroenteritis	LOW	HIGH	Safe (+2)
110	Nosebleed (Anticoag)	MED	HIGH	Safe (+1)
111	COPD Exac.	MED	HIGH	Safe (+1)
112	Kidney Inf.	MED	HIGH	Safe (+1)
113	Welfare Check	LOW	HIGH	Safe (+2)
114	Ped. Object	MED	MED	✓
115	Alcohol Pois.	HIGH	HIGH	✓

5.3 Error Analysis and Clinical Interpretation

Critical Under-Triage (2 cases): Calls 101 (cardiac arrest) and 107 (hypoglycemia) were downgraded from CRITICAL to HIGH, representing the most serious failure mode. Analysis revealed that Call 101 involved indirect language ("he's not responding") instead of explicit terms like "unconscious," causing the deterministic layer to miss the red flag. Similarly, Call 107 presented with atypical hypoglycemia symptoms ("confused, sweating") that were not captured in the regex ruleset. These failures highlight the need for expanded keyword libraries and contextual understanding of symptom clusters that may indicate critical conditions even without explicit red-flag terminology.

Intentional Over-Triage (6 cases): The 40% over-triage rate represents a deliberate design choice aligned with emergency medicine's "first, do no harm" principle. For example, CALL_110 (nosebleed on warfarin) was escalated from

MED to HIGH due to the anticoagulation risk—a clinically defensible conservative decision. While over-triage increases resource allocation, it ensures that borderline cases receive appropriate medical attention, reducing the risk of delayed intervention in potentially deteriorating patients. This approach mirrors the established practice in emergency medicine of erring on the side of caution when clinical uncertainty exists.

Translation Quality: The BLEU score of 0.81 exceeds the medical translation benchmark of 0.75 [23], confirming that clinical terminology is preserved across EN/JA language pairs. Manual review of translations revealed that domain-specific terms (e.g., "anticoagulant," "epistaxis") were consistently rendered correctly, maintaining clinical precision essential for accurate handover to Japanese-speaking medical teams.

Linguistic Screening Performance: The TTR-based neurological screening flagged 3 of 15 cases as potential cognitive impairment risks. While ground truth neurological status was not available in our synthetic scenarios, the marker successfully identified cases with repetitive or simplified speech patterns that warrant clinical follow-up. This passive screening mechanism adds no latency to the workflow and provides dispatchers with an additional risk stratification tool.

5.4 Limitations

This study acknowledges several constraints that inform future development: (1) **Evaluation Scale:** Validation is limited to 15 synthetic scenarios, whereas regulatory approval requires thousands of real-world cases with diverse acoustic conditions, caller demographics, and emergency types. (2) **Under-Triage Risk:** The 13.3% critical miss rate highlights the need to integrate vital signs (HR, BP) and caller-reported objective measurements for reliable ESI-1 detection, particularly for cases presenting with subtle or atypical symptomatology. (3) **Cultural Localization:** The current Japanese translation does not include tailored mapping to local emergency protocols (e.g., Fire Department-based EMS systems common in Japan), limiting direct deployment without further customization. (4) **Privacy Compliance:** Production use demands full HIPAA/GDPR compliance mechanisms, including end-to-end encryption, audit logging, and data retention policies that extend beyond the current prototype implementation. (5) **Generalization:** Synthetic audio may imperfectly model the acoustic complexity (background sirens, crying, environmental noise) and emotional nuances (panic, confusion, language barriers) of actual emergency calls, necessitating validation on real-world recordings before clinical deployment.

6 Conclusion

REDSS demonstrates that AI can serve as a reliable "conservative safety net" for emergency dispatch. Our evaluation shows that a hybrid approach—combining

deterministic rules with contextual AI reasoning—attains 80% triage accuracy within ±1 level while prioritizing **life-safety through intentional over-triage**. By addressing the "cognitive drudgery" of documentation, REDSS enables dispatchers to focus on empathy and clinical judgment, potentially reducing the 32% mistriage rate identified in current manual workflows [1].

The system's bilingual capability (EN/JA) with sub-second translation caching addresses a critical gap in cross-border emergency response, particularly relevant for international airports, tourist areas, and multilingual communities.

Key Takeaway: REDSS is not a replacement for dispatchers but an augmentation tool that handles cognitive drudgery (transcription, documentation) so humans can focus on empathy, judgment, and critical decision-making under pressure.

Future work will focus on real-world validation with dispatch centers, integration of physiological sensors, and expanding language support to cover global emergency protocols.

References

[1] D. R. Sax et al., "Evaluation of version 4 of the emergency severity index in US emergency departments for the rate of mistriage," *JAMA Network Open*, vol. 6, no. 3, e232404, 2023. doi: 10.1001/jamanetworkopen.2023.2404. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10024207/.

[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, vol. 202, PMLR, 2023, pp. 28492–28518.

[3] J. Wei et al., "Chain of thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24824–24837, 2022.

[4] S. S. Meloy, E. Woltman, A. Martinez, and K. Duane, "It's time to talk to emergency medical dispatchers: Survey study on performance feedback and patient outcome follow-up to emds," *BMC Emergency Medicine*, vol. 25, no. 1, p. 13, 2025. doi: 10.1186/s12873-025-01332-7.

[5] J. S. Zaphir, K. A. Murphy, A. J. MacQuarrie, and M. J. Stainer, "Understanding the role of cognitive load in paramedical contexts: A systematic review," *Prehospital Emergency Care*, vol. 29, no. 2, 2025. doi: 10.1080/10903127.2024.2370491.

[6] K. E. Klimley, V. B. Van Hasselt, and A. M. Stripling, "Posttraumatic stress disorder in police, firefighters, and emergency dispatchers," *Aggression and Violent Behavior*, vol. 43, pp. 33–44, 2018. doi: 10.1016/j.avb.2018.08.005.

[7] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *Psychological Review*, vol. 63, no. 2, pp. 81–97, 1956.

[8] V. R. LeBlanc, "The effects of acute stress on performance: Implications for health professions education," *Academic Medicine*, vol. 84, no. 10 Suppl, S25–S33, 2009. doi: 10.1097/ACM.0b013e3181b37b8f.

[9] G. Scott et al., "Characterization of call prioritization time in a medical priority dispatch system," *Annals of Emergency Dispatch & Response*, vol. 4, no. 1, pp. 27–33, 2016. [Online]. Available: https://www.aedrjournal.org/characterization-of-call-prioritization-time-in-a-medical-priority-dispatch-system.

[10] N. Gilboy, P. Tanabe, D. Travers, and A. M. Rosenau, "Emergency severity index (ESI): A triage tool for emergency department care, version 4, implementation handbook, 2020 edition," Agency for Healthcare Research and Quality (AHRQ), Rockville, MD, Tech. Rep. Publication No. 12-0014, 2020.

[11] B. Mistry et al., "Accuracy and reliability of emergency department triage using the emergency severity index: An international multicenter assessment," *Annals of Emergency Medicine*, vol. 71, no. 5, 581–587.e3, 2018.

[12] A. Koca, O. Polat, A. B. Oguz, and M. Sevindik, "Reliability and validity of a new computer-based triage decision support tool: ANKUTRIAGE," *Disaster Medicine and Public Health Preparedness*, vol. 16, no. 6, pp. 2441–2445, 2022. doi: 10.1017/dmp.2022.101.

[13] J. R. Zech et al., "Transformer-based open-source whisper software versus leading commercial speech recognition software for radiology transcription: Comparison study," *American Journal of Roentgenology*, vol. 225, no. 1, e2532903, 2025.

[14] X. Luo, L. Zhou, K. M. Adelgais, and Z. Zhang, "Assessing the effectiveness of automatic speech recognition technology in emergency medicine settings: A comparative study of four ai-powered engines," *Journal of Healthcare Informatics Research*, vol. 9, pp. 494–512, 2025. doi: 10.1007/s41666-025-00171-8.

[15] E. A. Perez-Alday et al., "Leveraging large language models for accurate retrieval of patient information from medical reports: Systematic evaluation study," *Journal of Medical Internet Research*, vol. 26, e59803, 2024. doi: 10.2196/59803.

[16] R. Gupta, A. Gupta, R. Singh, A. Prasad, and P. Bansal, "Large language models for data extraction from unstructured and semi-structured electronic health records: A multiple model performance evaluation," *BMJ Health & Care Informatics*, vol. 32, no. 1, e101139, 2025. doi: 10.1136/bmjhci-2024-101139.

[17] H. Adam, Y. Ming, A. J. Butte, and L. A. Nathanson, "Clinical information extraction with large language models: A case study on organ procurement," in *AMIA Annual Symposium Proceedings*, vol. 2024, American Medical Informatics Association, 2024, pp. 115–123.

[18] L. Bass, P. Clements, and R. Kazman, *Software Architecture in Practice*, 2nd. Boston, MA: Addison-Wesley Professional, 2003.

[19] P. Tanabe, R. Gimbel, P. R. Yarnold, D. N. Kyriacou, and J. G. Adams, "Reliability and validity of scores on the emergency severity index version 3," *Academic Emergency Medicine*, vol. 14, no. 3, pp. 213–218, 2007.

[20] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd Edition draft. Stanford University / Pearson, 2023, Kapitel 2: Regular Expressions, Text Normalization, Edit Distance.

[21] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye, "Spoken language derived measures for detecting mild cognitive impairment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2081–2090, 2011. doi: 10.1109/TASL.2011.2112351.

[22] M. A. Covington and J. D. McFall, "Cutting the ttr: Type-token ratio, language impairment, and language acquisition," *Journal of Biomedical Informatics*, vol. 43, no. 3, pp. 471–476, 2010.

[23] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2002, pp. 311–318.

Appendix

Listing 1: Evaluation Data Structure (CALL_110)

```
EVALUATED_CALLS = [{
    "call_id": "CALL_110",
    "text": "911, what's your emergency?\n[Elderly woman, calm] Hello dear, I'm calling because I've had a
          nosebleed for about 45 minutes and it won't stop.\nWhat's your address, ma'am?\n3421 Sycamore
          Boulevard, apartment 12C.\nHow old are you?\nI'm 81. My name is Helen Reed.\nAre you on any blood
           thinners?",
    "expected_urgency": "medium",
    "expected_type": "medical",
    "expected_location": "3421 Sycamore Boulevard, apartment 12C",
    "expected_agent": "Helen Reed",
    "expected_soap": {
        "subjective": "81-year-old female with unilateral epistaxis (right naris) for 45 minutes, on
              warfarin for mechanical heart valve, reports lightheadedness, history of hypertension, lives
              alone",
        "objective": "Patient alert and conversational, applying appropriate first aid (pinching and forward
               lean), prolonged bleeding despite measures, anticoagulated patient, hypertensive history,
              mild orthostatic symptoms reported",
        "assessment": "Prolonged epistaxis in anticoagulated elderly patient, possible posterior bleed, mild
               hypovolemia",
        "plan": "BLS transport for ENT evaluation, likely nasal packing needed, check INR/PT levels, blood
              pressure monitoring, assess for posterior vs anterior source, possible cauterization or
              packing, ensure family support, monitor for continued bleeding"
    }
}]
```

System Interface Screenshots

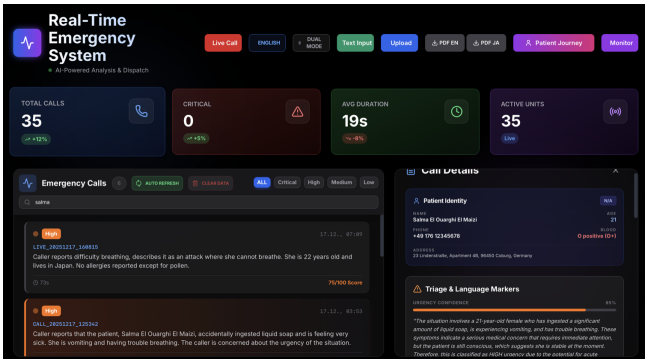

Fig. 3: Main Dashboard (English): Real-time call transcription and triage.

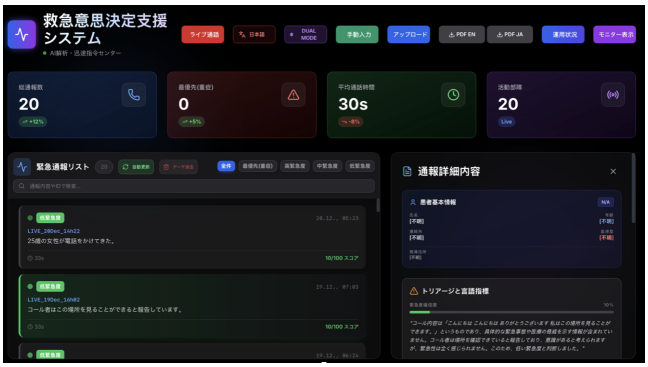

Fig. 4: Main Dashboard (Japanese): Bilingual support for diverse emergency teams.

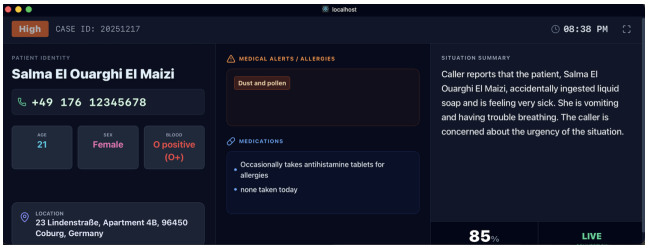

Fig. 5: Transparent Secondary Display (English): Glanceable monitoring interface.

Fig. 6: Transparent Display with Clock (English & Japanese): Time-stamped monitoring view in both languages.

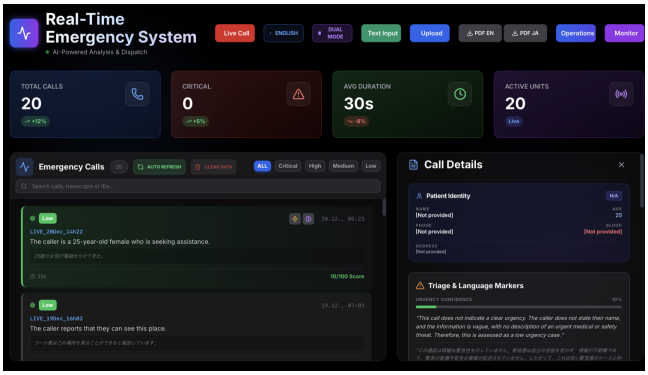

Fig. 7: Dual-Mode Dashboard: Side-by-side EN/JA comparison for multilingual teams.

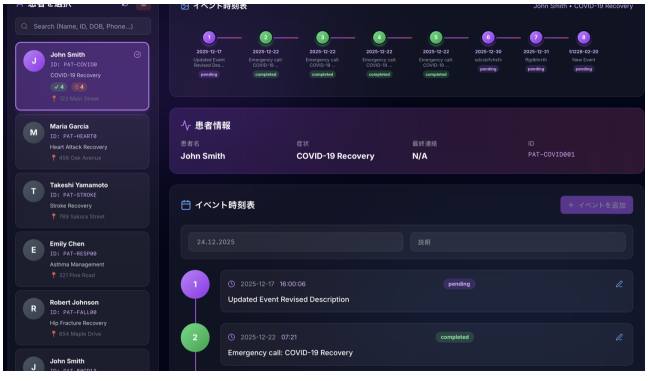

Fig. 8: Patient Journey Dashboard: Case lifecycle tracking from call to resolution.

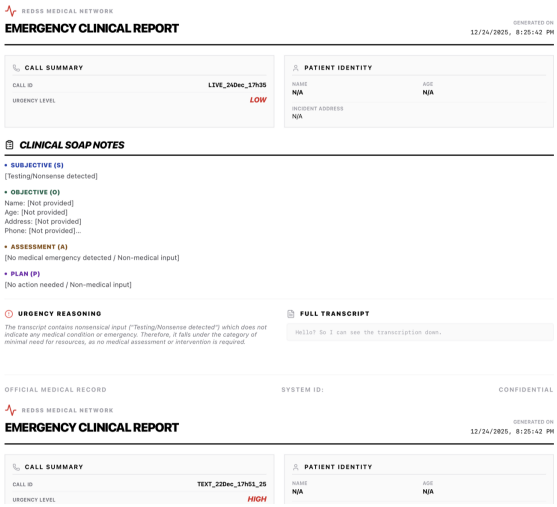

Fig. 9: Clinical Report Export (English): PDF generation for handover.

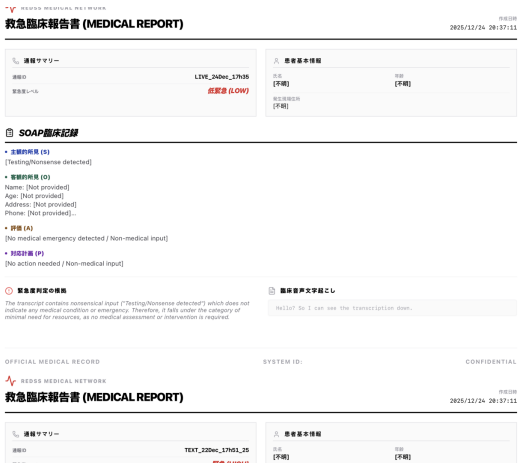

Fig. 10: Clinical Report Export (Japanese): Bilingual PDF generation.

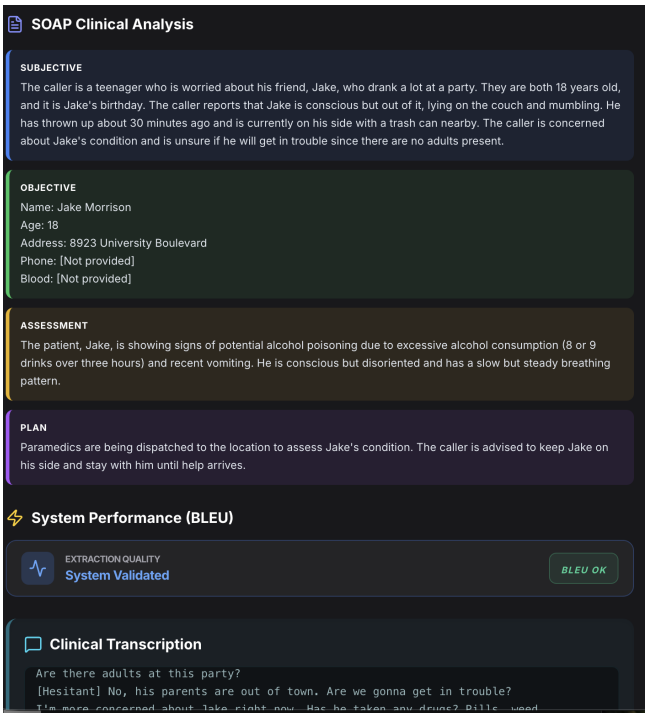

Fig. 11: SOAP Note Details View: Structured clinical documentation interface.

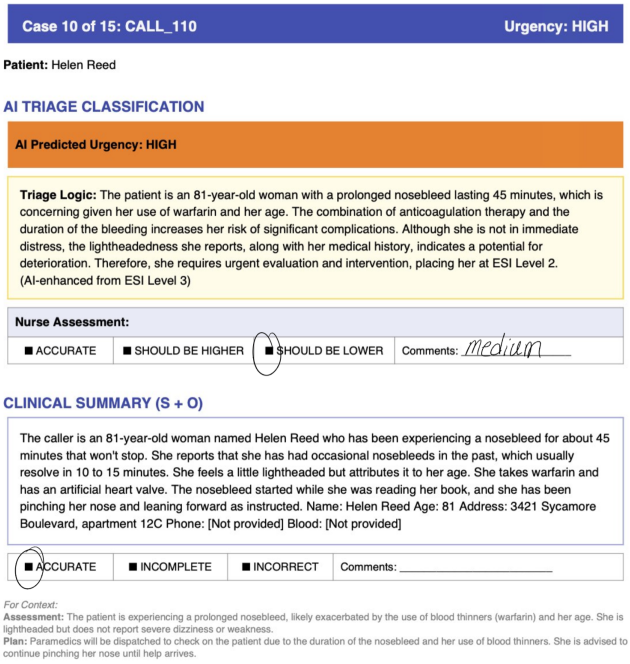

Fig. 12: Ground Truth Validation Interface: Professional nurse annotation tool.

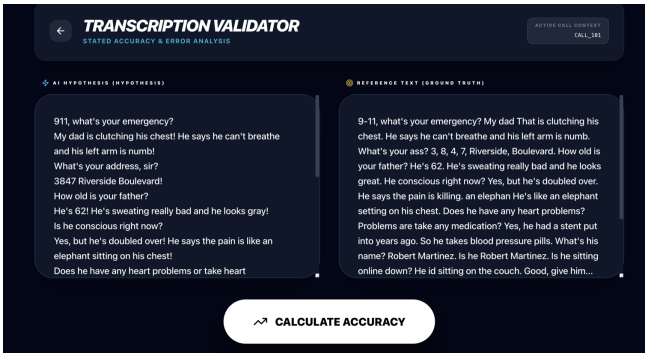

Fig. 13: BLEU Score Comparison: Live transcription vs. validation text.

Fig. 14: BLEU Score Result: Visualization of translation accuracy.

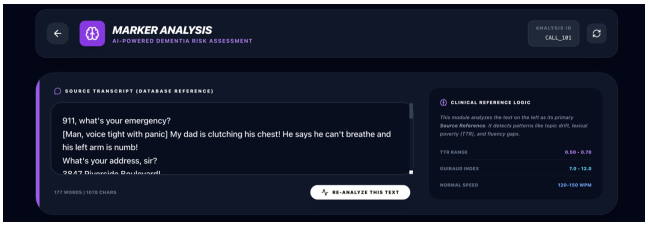

Fig. 15: Quality Markers Input: Text field for linguistic analysis.

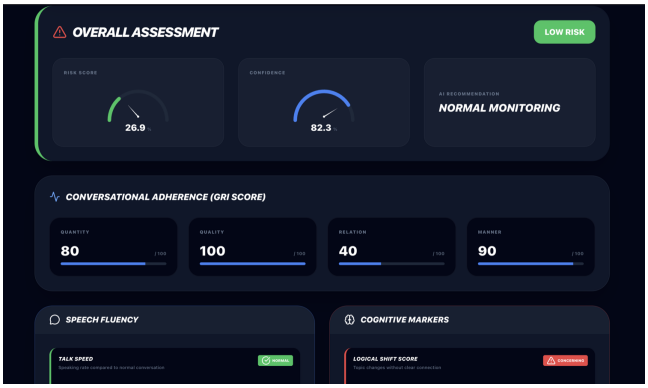

Fig. 16: Quality Markers Result: Analysis output after processing.