ARTICLE IN PRESS

npj Digital Medicine

https://doi.org/10.1038/s41746-026-02498-5

Article in Press

Deep language model-based early recognition of out-of-hospital cardiac arrest from real-time emergency calls

Received: 7 November 2025

Accepted: 19 February 2026

Cite this article as: Choi, H.-J., Hwang, M., Cho, S. *et al.* Deep language model-based early recognition of out-of-hospital cardiac arrest from real-time emergency calls. *npj Digit. Med.* (2026). https://doi.org/10.1038/s41746-026-02498-5

Hong-Jae Choi, Minyoung Hwang, Sangyeon Cho, Hyunsoo Kim, Junyeong Kim, Woori Bae & Changhee Lee

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

ARTICLE IN PRESS

Deep Language Model-based Early Recognition of Out-of-Hospital Cardiac Arrest from Real-Time Emergency Calls

Hong-Jae Choi[1†], Minyoung Hwang[2†], Sangyeon Cho[3,5], Hyunsoo Kim[4], Junyeong Kim[3], Woori Bae[6], Changhee Lee[2*]

[1]Department of Thoracic and Cardiovascular Surgery, SMG-SNU Boramae Medical Center, Seoul National University College of Medicine, Seoul, South Korea

[2]Department of Artificial Intelligence, Korea University, Seoul, South Korea

[3]Department of Artificial Intelligence, Chung-Ang University, Seoul, South Korea

[4]Incheon Fire Headquarters, Incheon, South Korea

[5]Korea Surgical Research Foundation, Seoul, South Korea

[6]Department of Emergency Medicine, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, South Korea

†These authors contributed equally to this work.

***Corresponding Author**

Changhee Lee, PhD

Assistant Professor,

Department of Artificial Intelligence, Korea University, Seoul, 02841, South Korea

Tel: +82-10-4105-2276

ARTICLE IN PRESS

ARTICLE IN PRESS

E-mail: changheelee@korea.ac.kr

ARTICLE IN PRESS

ARTICLE IN PRESS

Abstract

We developed a dynamic deep learning model (DyLM-OHCA) for early out-of-hospital cardiac arrest (OHCA) detection. Using 158,973 emergency call transcripts from three South Korean metropolitan regions, we trained DyLM-OHCA for 60 s OHCA identification and compared its performance against four conventional machine learning algorithms—Logistic Regression, XGBoost, Gradient Boosting, and Random Forest. DyLM-OHCA markedly outperformed all other benchmarks (AUROC=0.937; AUPRC=0.456). We analyzed global and sample-level word importance and temporally predicted OHCA risk patterns. Word attribution revealed differences in important words between callers and dispatchers. OHCA recognition was influenced more by conversational flow than by individual keywords. True-positive cases sustained high-risk scores, whereas over half of false-positive cases showed early risk score decline. DyLM-OHCA captures clinically meaningful dialogue patterns, moving beyond simple keyword spotting. By providing real-time, context-aware, and interpretable risk assessments, our model is potentially valuable in decision support, enhancing dispatcher confidence and improving early OHCA recognition.

ARTICLE IN PRESS

ARTICLE IN PRESS

Introduction

Out-of-hospital cardiac arrest (OHCA), a major global cause of death, occurs when the heart stops functioning and systemic circulation collapses outside a hospital setting[1]. Improving survival from OHCA depends heavily on early recognition and immediate initiation of cardiopulmonary resuscitation (CPR)[2,3]. For every minute of CPR delay, survival decreases by 7–10%, whereas when CPR is initiated promptly, this decline slows to 3–4% per minute[4]. Rapid bystander CPR can nearly double survival rates[4] and is recognized as a critical step in the OHCA chain of survival[5]. CPR within 3 min of OHCA onset can nearly triple survival, according to large-scale observational studies[6].

Despite the importance of early recognition, even experienced emergency dispatchers fail to identify OHCA in 20–30% of cases during emergency calls[7-9]. This recognition gap has been attributed to various factors, including incomplete or inaccurate information from callers, physical separation between the caller and the patient, premature call termination, and communication difficulties arising from emotional distress[10-12]. One critical barrier is the misinterpretation of agonal breathing—an abnormal breathing pattern often seen in cardiac arrest—as normal respiration, because callers often describe it as if the patient were breathing normally[10,11].

In contrast, when dispatchers successfully recognize OHCA during a call, patient outcomes improve significantly[7,13,14]. One study reported that when OHCA was recognized, the likelihood of bystander-performed CPR was 7.8 times higher, and the 30-day survival rate increased by a factor of 2.8 compared to that in unrecognized cases[7]. Thus, the recognition of OHCA is a key prerequisite for dispatcher-assisted CPR (DA-CPR)[3,5], and international guidelines recommend

ARTICLE IN PRESS

that dispatchers identify OHCA within the first minute of the call[3]. Timely and accurate OHCA recognition during the early moments of the call is essential to maximizing the chance of survival.

To address this issue, structured call interrogation protocols have long been used to assess consciousness and breathing[3,12]. However, these protocols are often misled by caller misinterpretation—especially when agonal breathing is mistaken for normal breathing[10,11]. To overcome these limitations, various strategies have been explored, including real-time audio analysis, closed circuit television-assisted triage, and, more recently, machine learning (ML) models[12]. A previous study developed an ML model trained on tens of thousands of emergency call recordings to support dispatcher decision-making[2]. These models demonstrated promising sensitivity and faster recognition of OHCA cases.

Notably, a randomized controlled trial showed that an ML system outperformed human dispatchers in identifying OHCA during live calls[15]. However, the improved predictive performance did not translate into significantly better recognition by dispatchers, largely because they did not consistently trust or act upon the system alerts[15]. This underscores a major limitation of current ML approaches; without contextual understanding or interpretability, their clinical utility remains constrained by user acceptance and workflow integration[15].

Recent advances in language models offer a promising alternative. Unlike conventional ML models that rely on manually engineered features, language models can directly process transcripts derived from emergency calls, retaining the conversational flow and context of the interaction[16-19]. This allows the model to capture context-dependent or vague expressions—such as those describing agonal breathing—that are often missed by traditional systems. Moreover,

ARTICLE IN PRESS

their ability to adapt to diverse linguistic patterns makes them particularly well-suited for real-time application in urgent clinical communications between callers and dispatchers[18,20].

This study aimed to develop and evaluate a dynamic deep language model (DLM)-based prediction model for the early recognition of OHCA (hereinafter referred to as DyLM-OHCA) using real-world emergency call data. Specifically, we assessed the ability of the model to identify OHCA cases within 60 s of call onset, and we analyzed key linguistic features that contributed to model predictions.

Results

Emergency call characteristics

The characteristics of emergency calls included in the study cohort are summarized in **Table 1**. Among a total of 158,973 emergency calls, the majority of the calls were classified as emergency medical services (EMS) calls (*n*=120,039; 75.5%), followed by search and rescue (*n*=22,817; 14.3%), firefighting (*n*=10,751; 6.8%), and miscellaneous emergency calls (*n*=5,366; 3.4%). Male callers represented 52.1% (*n* = 82,812) of total emergency calls.

In the EMS category, non-severe disease calls were most common (*n*=72,154; 45.4%), followed by trauma-related calls (*n*=16,903; 10.6%). OHCA cases accounted for 2,646 calls (1.7%). In addition, the distribution of patient symptoms documented at EMS arrival is presented in **Supplementary Figure 1**. Pain-related complaints were most frequently recorded, including other pain (n = 20,935, 12.7%) and abdominal pain (n = 12,602, 7.7%), followed by dizziness and general weakness.

The dataset was split using a stratified random sampling approach to preserve the proportion

ARTICLE IN PRESS

of OHCA cases between the training (*n*=127,178; 80.0%) and test (*n*=31,795; 20.0%) sets. No statistically significant differences were observed in emergency call characteristics between the two cohorts.

Performance of the prediction model

The performance of five OHCA recognition models in identifying cases within 60 s of call onset was evaluated (**Figure 1**). Overall, DLM-based approaches outperformed conventional ML models in terms of both the area under the receiver-operating characteristic curve (AUROC; **Figure 1a**) and the area under the precision-recall curve (AUPRC; **Figure 1b**). Among all models, DyLM-OHCA demonstrated the best performance (AUROC=0.937; AUPRC=0.456). Pairwise comparisons against the other models are summarized in **Supplementary Table 1**, and all tests showed $p < 0.001$. The performance of the models in OHCA recognition within 120 s is summarized in **Supplementary Table 2**, which showed modest gains in conventional ML models but no further improvement in the DyLM-OHCA.

To evaluate threshold-dependent performance, we assessed changes in accuracy, precision, specificity, F1-score, and Brier score at three levels of recall (true-positive [TP] rate): 0.75, 0.80, and 0.85 (**Supplementary Table 3**). Based on previous literature reporting dispatcher OHCA recognition rates between 70.0% and 80.0%, a recall value of 0.80 was selected as a clinically meaningful threshold for further analysis.

Table 2 presents all model performance metrics at the fixed recall level of 0.80. For the best-performing model (DyLM-OHCA), when the recall value was 0.805, precision reached 0.155, specificity was 0.924, and the corresponding false-positive (FP) rate (1 – specificity) was 7.6%.

ARTICLE IN PRESS

Table 3 presents the confusion matrix for DyLM-OHCA at a recall value of 0.80. DyLM-OHCA identified 426 TP OHCA cases and 2,324 FP cases in the test set.

Sensitivity analyses

Sensitivity analyses were conducted to assess model performance under alternative input restrictions related to protocol-related questions. Protocol activation was defined using the broad protocol definition, as the presence of either a consciousness-assessment or a breathing-assessment question.

First, performance was evaluated within the subset, comparing OHCA versus non-OHCA with severe disease. Broad protocol activation prevalence in these groups is shown in **Supplementary Table 4** (63.3% vs 56.0%). Model performance within this protocol-activated subset is reported in **Table 4**. Within this protocol-activated subset, DyLM-OHCA maintained high discrimination (AUROC = 0.904; AUPRC = 0.794). When protocol-related utterances were excluded while retaining the remaining dialogue, performance remained similar (AUROC = 0.888; AUPRC = 0.752).

Second, performance was assessed using caller-only text input in full cohort. Under the caller-only input setting, DyLM-OHCA retained meaningful discrimination (AUROC = 0.894; AUPRC = 0.394).

Third, performance was evaluated in the full cohort under the pre-protocol restriction. Compared with the full-input setting, discrimination decreased (AUROC = 0.857; AUPRC = 0.292). Similarly, within the protocol-activated subset, restricting the input to the pre-protocol segment also resulted in lower performance (AUROC = 0.770; AUPRC = 0.500). Among model-positive calls (TP+FP; n = 2,771), broad protocol-related questions were identified in 1,617 calls

ARTICLE IN PRESS

(**Supplementary Table 5**). Within these 1,617 protocol-present calls, first hitting time (FHT) occurred after protocol onset in 918 calls (56.8%) and before protocol onset in 699 calls (43.2%; **Supplementary Table 6). Supplementary Fig. 2** compares FHT time and protocol onset time across the 1,617 protocol-present model-positive calls. The mean time to FHT was 22.6 s (standard deviation [SD], 12.3), whereas the mean time to protocol onset was 26.5 s (SD, 12.5).

To complement the aggregate sensitivity analyses, sample-level risk-score trajectories with aligned dialogue excerpts are provided in **Supplementary Figures 3–5**. These examples include (i) a TP call with no protocol-related questions under the broad protocol definition (**Supplementary Figure 3**), (ii) a TP call in which FHT occurred before protocol onset (**Supplementary Figure 4**), and (iii) a FP call in which FHT also occurred before protocol onset (**Supplementary Figure 5**).

ARTICLE IN PRESS

Time to OHCA recognition

Figure 2a shows the distribution of OHCA recognition times using DyLM-OHCA. The median time to OHCA recognition was 19.7 s (interquartile range [IQR], 13.4–29.1), with a mean of 22.3 s (standard deviation [SD], 11.9). **Figure 2b** compares recognition times between TP and FP cases. For TP cases, the median recognition time was 16.7 s (IQR, 11.2–26.6), with a mean of 20.0 s (SD, 11.2). For FP cases, the median recognition time was 20.1 s (IQR, 13.4–29.7), and the mean was 22.8 s (SD, 11.9). This difference in mean recognition times between TP and FP cases was significant ($p<0.001$).

Word-level leave-out-one analysis at the FHT for OHCA detection

Fig. 3 summarizes word-level contributions estimated using a leave-out-one (LOO) analysis at

ARTICLE IN PRESS

the FHT, stratified by speaker type (caller vs. dispatcher) and prediction outcome (TP vs. FP). For each case, the LOO analysis was applied to words within the FHT utterance while the immediately preceding utterance was included in the model input to preserve conversational context.

Figure 3a illustrates the most influential words from caller utterances in TP cases. The words "looks like/think," "no," "don't/not," and "breathing" were among the most important, typically appearing in responses to dispatcher questions about the patient's consciousness or breathing status—for instance, "no" in response to "Is the patient conscious?" or "not breathing" when asked about respirations. A representative TP example in which "not" appears in the caller's response at the alert point is shown in **Figure 4**. Notably, "looks like/think" and "no" were consistently associated with OHCA recognition. While "don't/not" and "breathing" were important contributors, they also appeared in non-OHCA contexts. Additionally, personal information (e.g., detailed address) were found to play a meaningful role.

Figure 3b highlights the top-ranked words from dispatcher utterances in TP cases. "Conscious" emerged as the highest-ranked word among dispatcher utterances in TP cases, yet it was also frequently used in calls that were predicted to involve non-OHCA cases. Words such as "don't/not," "breath," and "is breathing" were frequently used in follow-up questions to clarify the patient's condition. Interestingly, "address" also appeared as a highly contributing word, suggesting that moments when the dispatcher confirmed the caller's location coincided with the OHCA predictions of DyLM-OHCA. A representative TP example in which FHT occurs during an "address"-related utterance is shown in **Supplementary Figure 3**.

Figure 3c shows the most important words from caller utterances in FP cases. Words like

ARTICLE IN PRESS

"breathing" and "don't/not," previously associated with TP predictions, also contributed in FP cases. In addition, words such as "someone" and "collapsed/fell down" were commonly used in descriptions like "someone collapsed" or "he's not waking up."

Figure 3d presents the top contributing words from dispatcher utterances in FP cases. "Passed away" was a highly weighted term, even though it was associated with FP predictions. Other words such as "conscious," "address," and "breath" also overlapped with those seen in TP dispatcher utterances, suggesting semantic ambiguity. The word "call/line" often appeared in statements like "please stay on the line," reflecting procedural language rather than clinical judgment. A representative FP example in which "conscious" appears in the dispatcher's questioning at the alert point is shown in **Supplementary Figure 6b**.

Figure 4 illustrates a sample-level prediction case demonstrating how the model recognized OHCA based on the conversational flow between the caller and dispatcher. The figure shows dynamic changes in OHCA risk score over time, mapped to corresponding utterances. At 10.7 s into the call, the caller stated, "He's not waking up, even when I call him," which triggered the initial OHCA recognition by the model. Following this utterance, the risk score continued to rise and remained elevated. A word-level LOO importance analysis identified that the phrases "not" and "waking up" were the most influential contributors to the decision of the model at that time.

Temporal patterns of OHCA recognition in TP and FP cases

Temporal patterns of OHCA prediction scores for TP and FP cases are shown in **Supplementary Figure 6**. **Supplementary Figure 6a** shows a TP case in which OHCA was first recognized at 10.7 s. The OHCA prediction score remained elevated throughout the rest of the call, indicating a sustained recognition pattern. In contrast, **Supplementary Figure 6b** shows an FP case in which

ARTICLE IN PRESS

OHCA was initially recognized at 11.5 s. However, as the conversation progressed, the prediction score declined below the threshold, suggesting a transient pattern of misrecognition. Although this call crossed the threshold and was labeled as OHCA at the FHT, subsequent dialogue led the model to revise its prediction.

Table 5 summarizes the distribution of TP and FP cases according to the persistence of the OHCA prediction score. Cases were categorized as sustained if the OHCA prediction score remained above the decision threshold after the FHT and as drop-off if the risk later declined below the threshold. Among the 2,771 OHCA-recognized cases, 426 were TPs and 2,345 were FPs. In the TP group, 397 cases (93.2%) showed a sustained prediction score, whereas only 29 cases (6.8%) showed a score that dropped below the threshold. In contrast, the FP group had 1,111 sustained cases (47.4%) and 1,234 drop-off cases (52.6%), indicating a significantly different temporal pattern between the groups (p<0.001). A synthesis of representative dialogue patterns at the alert points and their operational pathways is provided in **Box 1**, with corresponding distributions shown in **Supplementary Table 7**. These representative patterns were summarized by focusing on the model's high-confidence trigger utterances at the first alert (FHT) and organizing them by pathway.

ARTICLE IN PRESS

Discussion

In this study, we developed a language model to support the early recognition of OHCA using real-world emergency call data. The model showed the potential to identify OHCA cases within 60 s from call onset, suggesting its usefulness for timely and actionable predictions. To improve interpretability, we applied a word-level LOO approach to examine key linguistic features and

ARTICLE IN PRESS

contextual cues that contributed to the model output. These findings offer insights into how OHCA may be recognized in natural emergency dialogues, how quickly it can be detected using our DLM-based prediction model, and how such models could assist dispatcher decision-making by providing both predictive performance and interpretability.

DyLM-OHCA is inherently context-dependent; therefore, the word-level LOO results are intended only as supportive evidence to help describe model prediction, rather than as definitive identification of "key words" for OHCA recognition. In our analysis, OHCA recognition appeared to occur not through a single keyword, but through the progressing dialogue between caller and dispatcher—specifically when the dispatcher inquired about the patient's consciousness and the caller responded that the patient was not waking up. This case illustrates how recognition may emerge from the broader conversational context, rather than individual words alone, suggesting that interactional structure can play a meaningful role in supporting OHCA detection. While previous studies have suggested that both caller and dispatcher speech may play a role in recognizing OHCA[21,22], much of the prior work has tended to focus on the caller's descriptions—such as reported symptoms, environmental cues, or basic patient information—often through keyword-based approaches[7,21,23,24]. However, emergency calls naturally involve interactive dialogue, and our analysis indicates that dispatcher speech—particularly follow-up questions and confirmation requests—was also associated with OHCA recognition. These expressions, which may reflect clinical concern or an effort to clarify uncertain information, appeared to contribute meaningful contextual cues that supported the decision-making process of the model. This aligns with prior suggestions that timely and appropriate follow-up questions by dispatchers—particularly concerning abnormal or absent

ARTICLE IN PRESS

breathing—may facilitate earlier OHCA recognition, potentially reducing delays[21,24].

Notably, we observed differences in caller- and dispatcher-derived feature contributions to OHCA recognition across TP and FP cases. Among caller-derived features, some of the highest-ranked words in TP cases included expressions such as “looks like/think” and “no.” Interestingly, this pattern contrasts with that reported in prior studies that emphasized more direct symptom descriptions[7,21,23,24]. In our analysis, these words were frequently used in response to dispatcher questions. For example, “no” was often used to answer inquiries about consciousness (e.g., “Is he awake?” can lead to a response “No”), whereas phrases like “I think he’s not breathing” appeared in response to breathing-related questions. This highlights how indirect or subjective responses—rather than concrete symptom terms—can still carry strong predictive value when embedded within the interactive context of a call. However, caller-derived words such as “breathing,” “conscious,” “collapsed,” and “don’t/not” also showed high importance in FP cases. This suggests that caller descriptions of patient status may not always be reliable, a finding that is consistent with prior literature[12]. Although these words are linguistically prominent, they may reflect inaccurate information, thereby leading to FP OHCA classifications.

From the dispatcher’s side, key TP-associated expressions included medically oriented questions involving “conscious” and “breathing,” reflecting clinically relevant probing[25]. Interestingly, we also found that in FP cases, dispatcher follow-up words such as “passed away?”—typically reflecting efforts to clarify the caller’s report of the patient being deceased—were associated with OHCA recognition. This aligns with prior research suggesting that when dispatchers re-confirm callers’ declarations of death, the process can facilitate early OHCA recognition, but it may also reflect potentially inaccurate assessments of patient viability[12],

ARTICLE IN PRESS

requiring careful interpretation and follow-up during the call. As shown by Riou et al., although callers who state that the patient is "dead" are more likely to decline from performing CPR, some patients still achieve return of spontaneous circulation, emphasizing that such statements should not lead to premature OHCA exclusion[26].

While individual keywords such as "breathing" or "consciousness" contributed to OHCA recognition, predictions by DyLM-OHCA were more strongly influenced by how these words were situated within the broader conversational context. Consistent with this, we summarized representative dialogue trigger patterns at the model's first-alert (FHT) and practical call-taking pathways (caller-led, dispatcher-led, and workflow/context-led), along with their distinct post-alert trajectories (**Box 1**). This finding highlights the importance of contextual interpretation over isolated word-level features in emergency calls. Unlike traditional models that often rely on fixed keyword matching[2,8,21,27], DyLM-OHCA recognized OHCA by incorporating conversational context. Rather than responding to single terms, our model interpreted meaning through the flow of the dialogue, thereby detecting critical moments based on the accumulation of information over time[28]. This context-sensitive approach highlights a key difference between our model and previous OHCA detection models, and thus may help overcome a common limitation reported in previous randomized controlled trials: low dispatcher adoption due to a lack of model transparency[15]. By offering explainable attribution cues consistent with real conversational flow, DyLM-OHCA may enhance dispatcher confidence and serve not as a replacement for clinical judgment but as a supportive tool that reinforces decision-making in time-critical scenarios[29].

DyLM-OHCA achieved a median OHCA recognition time of 19.7 s, with approximately 75% of OHCA-positive calls identified within the first 29.1 s. At the 60th second, DyLM-OHCA

ARTICLE IN PRESS

maintained a specificity of 92.4%, a FP rate of 7.6%, and a positive predictive value (PPV) of 15.5%. Globally, dispatcher-based OHCA recognition rates typically range from 70–80%[7-9]. However, these results generally reflect OHCA recognition at the end of the call rather than within the first minute. Although international guidelines recommend OHCA identification within the first minute of the call[3], only a limited number of studies have evaluated model performance specifically within this timeframe. One such report found that the 60 s recognition rate was only 25–36%[27].

Blomberg et al. reported on an ML model with an overall OHCA recognition rate of 84.1% and specificity of 97.3%, but the median time to recognition was 44 s[2]. Similarly, Byrsell et al. reported OHCA recognition rates ranging from 81–90%, with specificity between 97–99% depending on the sensitivity threshold[27]. However, the time to recognition was considerably longer (median 52–85 s), and the 60 s recognition rate dropped to approximately 27–50%[27].

Compared to the models in these prior studies, our model demonstrated significantly faster recognition, reducing the median time by over 20 s while maintaining acceptable sensitivity and specificity. This early recognition capability may provide dispatchers with a crucial time window in which they can ask targeted follow-up questions and act decisively to initiate DA-CPR within the critical first minute. Such real-time risk signaling may support earlier clarification and timely initiation of DA-CPR when appropriate, although prospective dispatcher-in-the-loop evaluation will be needed to determine its impact on clinical outcomes.

Although the PPV of our model (15.5%) was lower than that reported in previous studies (approximately 33% at the end of call)[2], this finding should be interpreted cautiously and may imply a non-trivial false-alarm rate in operational use. In this context, we note that prior work

ARTICLE IN PRESS

has reported that a substantial proportion of FP OHCA alerts resulted in hospital transport (up to 87.0%)[2]; however, this observation does not negate the potential resource implications of false alerts. In our cohort, 22.4% of FP predictions occurred in patients ultimately classified as severe but non-OHCA conditions (e.g., Canadian Triage and Acuity Scale [CTAS] level 1–2), suggesting that some false OHCA predictions may co-occur with other time-sensitive emergencies. Nevertheless, these cases remain FP with respect to the primary target condition as OHCA, and further prospective evaluation is needed to quantify the operational burden, downstream actions, and net clinical impact of integrating real-time alerts into dispatch workflows.

In addition, among the FP cases, 22.4% (n = 524) were classified as severe but non-OHCA presentations. Within these 524 cases, the reported on-scene symptoms included unconsciousness (n = 218), shortness of breath (n = 115), seizures (n = 62), and syncope (n = 51) (**Supplementary Table 8**). These observations indicate that a subset of FP alerts occurred in high-acuity clinical contexts. However, these cases remain FP with respect to OHCA, and the extent to which such alerts are actionable or impose additional operational burden should be evaluated in prospective workflow studies.

Furthermore, 52.6% of the FP cases exhibited a decrease in OHCA prediction score during the continuation of the call, suggesting that these predictions reflected dynamic changes in real-time OHCA assessment. These findings indicate that FP cases were more likely to show a decline in prediction score following initial OHCA recognition. In real-world settings, such temporal patterns could be used by dispatchers to reevaluate the situation and reduce the impact of FP cases on clinical decision-making.

ARTICLE IN PRESS

Therefore, these findings support the clinical acceptability of the PPV of DyLM-OHCA as a trade-off for early and actionable intervention in time-sensitive emergencies situations. Importantly, the ability of DyLM-OHCA to recognize OHCA within 60 s—while maintaining a favorable balance of sensitivity and specificity—demonstrates its potential as a real-time decision-making support tool for emergency dispatchers.

This study presents two key methodological contributions to the early recognition of OHCA using real-world emergency call data. First, unlike conventional ML models that rely on predefined features and static word associations[2,15,27], our approach utilized a DLM capable of interpreting the dynamic flow of conversation between callers and dispatchers. This allowed our model to recognize OHCA not solely based on isolated keywords but through accumulated contextual understanding over time[16-19]. Hence, our proposed DyLM-OHCA model reflected how humans interpret clinical information in real-time, thereby enabling more flexible and adaptive predictions across diverse dialogue structures.

Second, beyond model performance metrics, we qualitatively examined how DyLM-OHCA arrived at its OHCA recognition decisions using several interpretability techniques, including global and sample-level word attribution, word-level importance scoring[30,31], and temporal prediction score patterns. These analyses demonstrated that the model recognized OHCA not merely through isolated keywords, but by capturing contextual meaning within the conversational flow, highlighting the importance of discourse-level interpretation. Such transparency enhances trust in the model outputs and supports its implementation in emergency call contexts where explainability is critical.

Therefore, these contributions suggest that DyLM-OHCA can improve not only the

ARTICLE IN PRESS

accuracy but also the interpretability of OHCA recognition systems. By addressing the issue of model opacity—previously cited as a barrier to adoption in randomized controlled trials[15]—this approach may help overcome resistance from emergency medical dispatchers and promote real-world clinical integration.

There are some limitations to this study. First, DyLM-OHCA was developed and evaluated using retrospective emergency call data from a single country, and the reported performance therefore reflects internal validation within the same national and operational context. As a result, the current findings may overestimate effectiveness when the model is deployed in different EMS systems, languages, or call-taking workflows. In addition, because the model was trained exclusively on Korean conversations, its applicability is presently limited to Korean-language settings. Future studies should include prospective evaluation and external validation across diverse EMS infrastructures and languages; within Korea, temporal validation using future emergency call data will also be needed to assess robustness over time. Second, the content and structure of dispatcher–caller dialogue may vary by EMS protocols, system interfaces, and dispatcher experience. Such differences could influence the linguistic patterns available to the model and affect transferability across settings. Third, because the model relies on unstructured text inputs, performance may be sensitive to transcript quality. Variability in speech-to-text accuracy, annotation conventions, or recording conditions may introduce noise and affect prediction reliability. Finally, this study did not evaluate real-world integration, usability, or the behavioral impact of model outputs on dispatcher decision-making. In addition, because dispatcher recognition labels and decision times were not available, the incremental utility of DyLM-OHCA in dispatcher-missed or delayed OHCA cases could not be quantified;

ARTICLE IN PRESS

such evaluation will require prospective dispatcher-in-the-loop studies or datasets with explicit dispatcher decision annotations. In operational settings, simply presenting an alert may be insufficient to change decisions under cognitive and workflow constraints; therefore, DyLM-OHCA should be considered a real-time, interaction-aware decision-support tool that provides an additional risk signal during the evolving dispatcher–caller dialogue rather than an automated decision-maker. Prospective simulation and field evaluations will be needed to assess feasibility, human factors, and whether the alert can support structured clarification when uncertainty remains.

In conclusion, we developed a dynamic DLM-based system to recognize OHCA in real time within the first minute of emergency calls. In retrospective evaluation, DyLM-OHCA demonstrated strong discrimination and rapid alerting. Qualitative examination of model behavior suggested that OHCA recognition is driven more by evolving conversational context across turns than by any single keyword. By providing an early risk signal during an ongoing call, DyLM-OHCA may support dispatcher recognition and decision-making when uncertainty remains.

ARTICLE IN PRESS

Methods

Study design and setting

This retrospective cohort study primarily adhered to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis + Artificial Intelligence checklist to ensure methodological rigor and transparency in reporting the language model-based prediction model[32] (**Supplementary Table 9**).

ARTICLE IN PRESS

In South Korea, EMS are provided through a national, fire-based, single-tiered system. Emergency calls are routed through 119, and each of the 17 provinces operates one or more emergency dispatch centers. When a call involves a suspected medical emergency such as cardiac arrest, it is transferred from a call-taker to a trained medical dispatcher. The dispatcher follows a standardized protocol, beginning with two critical questions: “Is the patient conscious?” and “Is the patient breathing normally?” If both answers are “no,” DA-CPR is initiated according to national guidelines[33].

Data source and study sample

This study used data from the AI Hub database, a national artificial intelligence infrastructure operated by the Ministry of Science and ICT and the National Information Society Agency of South Korea. The dataset is conditionally accessible from the AI Hub platform (https://www.aihub.or.kr) following user registration and agreement to the terms of use.

This study utilized the 119 Intelligent Emergency Call Voice Recognition Dataset, a comprehensive collection of 158,973 emergency call records (**Figure 5**). These data were collected in 2023 from fire departments in three metropolitan regions of South Korea: Seoul (n=104,496), Incheon (n=28,428), and Gwangju (n=26,049). The total duration of the recorded emergency call audio amounted to 3,064 h. Given the scale of the corpus (158,973 calls; 3,064 h), the available sample size was considered sufficient for model development and evaluation, in line with prior methodological guidance[34].

Data collection and definition

ARTICLE IN PRESS

Each record in the dataset comprised two primary components: a high-fidelity textual transcription and structured metadata. The transcriptions were initially generated using an automated speech-to-text engine and were subsequently manually reviewed and corrected by human annotators to ensure accuracy. The corresponding metadata, provided in JSON format, was annotated at both the case and utterance levels. Case-level metadata included variables such as caller gender, the primary reported symptom, and incident categorization. Utterance-level metadata provided a granular segmentation of the conversation, detailing the speaker role (e.g., dispatcher, caller, and third party), precise start and end timestamps for each utterance, and the aligned text segment from the full transcription.

The above transcription and curation procedures were conducted during the AI-Hub dataset construction process prior to our access. In this study, we used the released transcripts and metadata as provided, without additional manual correction or study-specific transcript cleaning. According to the AI-Hub documentation, calls with a duration of < 30 seconds and duplicate calls were excluded during dataset curation. Because we did not operate the speech-to-text pipeline and were not provided with paired audio–transcript benchmarks, we could not quantify the word error rate within the scope of this study and therefore evaluated model performance on the released transcripts as available to downstream users. Case-level metadata were used for descriptive characterization of the dataset and were not included as model input features. Dispatcher-level diagnostic decisions (e.g., whether/when the dispatcher suspected OHCA) were not recorded in the dataset, and no dispatcher decision timestamps were available. Instead, the only outcome labels available were ground-truth case labels derived from post-dispatch EMS documentation (i.e., field records). Therefore, analyses comparing DyLM-OHCA against

ARTICLE IN PRESS

dispatcher recognition—such as identifying dispatcher false-negative OHCA cases and estimating timing advantage—could not be performed in this study. Accordingly, throughout this manuscript, TP/FP/FN/TN and "false positive" refer to model predictions relative to the ground-truth case label, not to dispatcher recognition.

Severe disease without OHCA cases were defined as emergent clinical conditions requiring immediate intervention or transport, such as unstable vital signs[35], but without evidence of cardiac arrest (CTAS level 1–2).

To characterize the temporal stability of the model output after the initial alert, we additionally categorized model-positive calls based on the post-alert trajectory of the risk score. A sustained case was defined as a call in which, after the FHT, the predicted risk score never fell below the predefined threshold at any subsequent time point. A drop-off case was defined as a call in which, after FHT, the predicted risk score dropped below the threshold at least once. This classification was used to describe whether the model's elevated risk signal remained stable following the first threshold crossing.

ARTICLE IN PRESS

Outcomes

The primary outcome for this study was the correct recognition of OHCA by the model within the first 60 s of the call. Secondary outcomes included (i) recognition of OHCA within a 120 s timeframe; (ii) the overall time-to-recognition for TP cases; and (iii) the identification of critical keywords and contextual cues that contributed to OHCA recognition by the model. OHCA was defined from the official case report, which was completed by the on-scene EMS providers. This report contained the definitive on-site assessment of the patient's condition and served as the

ARTICLE IN PRESS

gold-standard reference for model training and evaluation.

Model development

For model development, we randomly partitioned the dataset into a training set (80.0% of the population, n = 127,178) and a held-out test set (20.0%, n = 31,795). The training set was further divided into 80.0% (n = 101,742) for model training and 20.0% (n = 25,436) for validation, which was used for hyper-parameter tuning and early stopping. To develop the ML-based prediction models, we first identified the most frequent terms by selecting the top 776 words, which together accounted for 85.0% of the total word frequency. These features were quantified using the term frequency-inverse document frequency (TF-IDF) scores[36]. Four widely used ML models were trained on these TF-IDF features: Logistic Regression, Gradient Boosting, Random Forest (all implemented with scikit-learn v1.4.0), and XGBoost (implemented with xgboost package v1.7.6). Hyper-parameters were optimized through a comprehensive grid search. For Logistic Regression, we evaluated 100 regularization coefficients logarithmically spaced between 10e-4 to 10e4. For the ensemble tree models (Random Forest, Gradient Boosting, and XGBoost), we explored maximum tree depths ranging from 1 to 5 and varied the number of estimators between 100 and 500. For the DyLM-OHCA model, to prevent overfitting and ensure optimal performance, we selected the model with the lowest validation loss, using a batch size of 12 and applying early stopping in the fine-tuning phase. The learning rate for fine-tuning was set to 10e-5 for both the GPT backbone and the classifier.

Advantages of Deep Language Models over Traditional Approaches

ARTICLE IN PRESS

ARTICLE IN PRESS

Traditional machine learning approaches for text classification, such as topic models and classical supervised learning methods, fundamentally operate on bag-of-words representations. These approaches treat text as an unordered collection of words, relying solely on word frequency and co-occurrence patterns without considering word order or sequential context. Consequently, they cannot distinguish between phrases where word order critically changes meaning.

Our transformer-based model processes emergency call text sequentially through self-attention mechanisms[37], which evaluate each word in relation to the entire conversational context rather than treating words as independent features[38]. This enables the model to distinguish between similar phrases based on surrounding information and capture multi-word dependencies that extend beyond simple word proximity. For instance, while "he's breathing" might indicate normal respiration in isolation, the model learns to recognize OHCA when this phrase appears alongside contextual cues such as "unresponsive," "collapsed suddenly," or breathing modifiers like "weird" or "gasping"—identifying patterns where multiple contextual elements collectively suggest OHCA even when no single phrase explicitly indicates cardiac arrest. Through training on 119 call transcripts with confirmed OHCA outcomes, the model learned associations between these context-dependent expression patterns and cardiac arrest, enabling recognition of cases that keyword-based systems would miss.

Moreover, traditional ML models are ill-suited for real-time emergency call processing. Since these models treat the entire text as a fixed input, they must reprocess the complete conversation from the beginning each time a new utterance is added. This becomes computationally expensive and introduces processing delays as the conversation progresses. In

ARTICLE IN PRESS

contrast, our decoder-based transformer architecture processes emergency calls incrementally[38]. When a new utterance is added to an ongoing conversation, the model retains the previously computed representations and only processes the new information, enabling efficient real-time OHCA detection as the call unfolds.

Dynamic risk prediction for early OHCA detection

This study proposes a novel framework that leverages DyLM-OHCA for dynamic, real-time OHCA prediction from emergency call transcripts (**Figure 6**). A key feature of our approach is its ability to train using only call-level diagnostic labels while producing continuous risk assessments as the call unfolds. This is achieved through a three-stage development process designed for real-time application: (i) weakly supervised training to generate dynamic token-level predictions, (ii) post-processing to aggregate and stabilize these raw predictions into sentence-level risk trajectories, and (iii) actionable alerting via optimized thresholding.

To enable dynamic risk prediction using only call-level labels, we employed a weakly supervised learning framework. As illustrated in the architecture of **Figure 6c**, the proposed DLM-based prediction model is built upon a GPT-2 backbone[38], a decoder-only model whose inherent causal attention mechanism effectively captures the temporal dynamics of conversations. The decoder structure processes the conversation sequentially, token by token, allowing the model to incorporate the full context of all preceding tokens as new conversational inputs (i.e., tokens) are dynamically added. Crucially, as the model reads a new token, it inherently considers all previously presented conversational content. This allows our model to generate a contextual representation at each token that cumulatively encapsulates all prior

ARTICLE IN PRESS

information from the start of the call. This process of sequential unfolding creates progressively informed representations, which guide the model to learn the temporal context of an escalating OHCA emergency.

A single, shared prediction head (i.e., a non-linear projection layer) is then applied to every token representation to produce token-level logits. The model is trained end-to-end using a multi-class cross-entropy loss, where all token-level predictions are supervised by the single call-level label. This objective requires the model to discriminate between three classes (i.e., OHCA, severe cases without OHCA, and non-severe cases), thereby compelling it to learn a more discriminative feature space than a simple binary classification (i.e., OHCA and non-OHCA cases) would allow. The empirical comparison between the two-way and three-way training objectives, including OHCA detection performance (AUROC/AUPRC) and false-alert characteristics (total FP and the proportion of severe non-OHCA among FP), is reported in **Supplementary Table 10**.

This end-to-end design is highly efficient for real-time inference. By inherently generating cumulative representations, it bypasses the significant computational overhead and latency of traditional methods that require iterative feature recalculation (e.g., TF-IDF), making it a practical solution for time-critical applications.

The raw token-level risk scores generated by the model can fluctuate during a call, which may occasionally lead to unwanted false alarms if used directly in a clinical setting. To ensure stable and reliable early detection, we implemented a two-step smoothing procedure as depicted in **Figure 6a**.

In the first step of our smoothing procedure, we aggregated the raw token-level risk scores

ARTICLE IN PRESS

produced by the model into sentence-level risk scores (S_n). Specifically, we extracted only the probabilities corresponding to the OHCA class predictions (L1 in **Figure 6c**). Since token-level risk scores—the smallest textual prediction units—may contain noise from non-informative or irregular tokens, we computed the average of these scores across all tokens within each sentence. By averaging at the sentence level, we transitioned the risk score unit from tokens to a more semantically coherent sentence representation, reducing noise and providing a stable base for further analysis. In the second step of our smoothing procedure, we applied the moving average (MA) to the sequence of sentence-level risk scores (S_n) to further smooth abrupt, non-informative fluctuations. The MA computes the mean of sentence-level risk scores within a fixed window of size *K* with striding 1 (which shifts forward by one sentence at a time) as $\tilde{S}_n = \frac{1}{K}\sum_{i=n-K+1}^{n} S_i$. In our experiments, we set $K = 3$. Together, these two steps produce a smoothed and reliable sentence-level prediction that captures both local stability and contextual dynamics.

This MA calculation ensures that the final post-processed risk score ($\tilde{S}_n$) for a given sentence reflects not only its immediate score but also the short-term risk trend established by the most recent K sentences. Consequently, this two-step process ensures that the post-processed risk score for each sentence reflects both its own risk score and the trend of recent sentences, effectively dampening transient fluctuations and improving the reliability of the final prediction.

To operationalize the model, we performed a cumulative analysis on the post-processed risk scores to identify the earliest possible detection point—the moment the score exceeds a pre-determined activation threshold. This threshold was carefully calibrated on a validation set by analyzing the precision-recall trade-off to satisfy a key performance requirement: achieving at least 80% recall for decisions made within 60 s. In real-time emergency call handling, our model

ARTICLE IN PRESS

makes a definitive OHCA decision as soon as the risk score crosses this boundary (**Figure 6b**). This allows the dispatch process to be initiated immediately, significantly reducing the time to intervention compared to a workflow that requires call completion.

Assessment of model performance

We evaluated the performance of the proposed model through both quantitative metrics and qualitative analysis. To align with the operational requirement of recognizing OHCA within 60 s, all quantitative evaluations and threshold optimizations were performed on call data truncated to the first 60 s. This ensures that the reported performance metrics reflect the utility of the model in a real-time decision-making scenario. To ensure statistical robustness, all performance metrics are reported as the mean and SD across 10 bootstrap samples, each with the same size as the test set. For conventional ML models, we truncated the input transcripts at the specified time points (60 s and 120 s from call onset) and applied feature extraction (TF-IDF) to these time-limited segments.

We evaluated quantitative performance using a comprehensive set of metrics. Overall classification was assessed with the AUROC, AUPRC (to account for class imbalance), Brier score (for calibration), accuracy, F1 score, precision, and recall. To capture early detection capability, we measured the FHT, defined as the elapsed time until the model prediction score first exceeds a predefined threshold in TP cases. FHT was computed using the cumulative analysis framework described above under the *"Actionable alerting via optimized thresholding"* subsection, with thresholds optimized in the validation set.

The ability of the model to update predictions dynamically in real time was evaluated by

ARTICLE IN PRESS

examining the temporal evolution of its risk scores in the test set. We first visualized the discriminative power of the model by plotting the average prediction score over time, stratified by outcome (OHCA and Non-OHCA), to highlight the divergence in risk trajectories between the two groups. Furthermore, we assessed the stability of the predictive signal after an initial detection at the FHT. By comparing the post-detection signal persistence in TP versus FP cases, this analysis validates the reliability of the early-stage decisions of the model.

To elucidate the rationale behind the model predictions, we conducted a two-part qualitative analysis. First, we performed a case study review of representative TP, FP, and false-negative predictions. For each case, the call transcript was juxtaposed with the corresponding risk trajectory of the model to provide insight into its predictive behavior. Second, to quantify the contribution of individual keywords to a positive detection, we employed a word-level LOO analysis[30,31]. This was performed on the trigger sentence, defined as the utterance corresponding to the FHT. While preserving the preceding conversational context, we systematically removed each word from the trigger sentence and measured the resultant drop in prediction probability, using the magnitude of this drop as a proxy for the importance of the word.

ARTICLE IN PRESS

Statistical analysis

Continuous variables are summarized as means with SDs or medians with IQR, whereas categorical variables are presented as frequencies and percentages. Group comparisons were performed using independent *t*-tests for continuous variables and chi-squared tests for categorical variables. DeLong's test was used to compare AUROC between DyLM-OHCA and the other models and to assess whether the observed performance differences were statistically

ARTICLE IN PRESS

significant[40]. All statistical analyses were performed using Python version 3.9.18 (Python Software Foundation, Beaverton, OR, USA). All *p*-values were two-tailed, with statistical significance defined as $p<0.05$. Model performance metrics were reported as the mean and SD across 10 independent repetitions for each experimental setup.

Ethical approval

The study was conducted according to the ethical principles outlined in the Declaration of Helsinki and approved by the institutional review board of Chung-Ang University (IRB No. 1041078-20230202-HR-031). Owing to the retrospective nature of the study, the requirement for informed consent was waived by the ethics committee of Chung-Ang University. All data were anonymous and de-personalized.

ARTICLE IN PRESS

Data Availability

The dataset used in this study is not publicly available due to the access policies and approval requirements of the providing organization. Specifically, the dataset is sourced from the AI Hub Public Data Platform (South Korea) and is designated as “The 119 Intelligent Emergency Call Voice Recognition dataset”. Access to the data can be obtained by following the official application and approval process on the AI Hub platform (https://www.aihub.or.kr/).

Code Availability

The code underlying this study is publicly available at the DyLM-OHCA repository and can be accessed at https://github.com/ggomaeng514/DyLM-OHCA.

ARTICLE IN PRESS

Acknowledgements

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00358602), Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) Artificial Intelligence Graduate School Program (No. RS-2019-II190079 [Korea University] and RS-2021-II211341 [Chung-Ang University]), and Artificial Intelligence Star Fellowship Support Program to nurture the best talents (RS-2025-02304828). The funder played no role in the study design; data collection, analysis, and interpretation; or manuscript writing.

Author Contributions

H-JC: Conceptualization, formal analysis, investigation, writing-original draft, and writing-review and editing. MH: Methodology, data curation, formal analysis, investigation, writing-original draft, and writing-review and editing. SC: Methodology, data curation, formal analysis, investigation, and writing-review and editing. HK: Investigation, data curation, and writing-review and Editing. JK: Methodology, data curation, formal analysis, investigation, and writing-review and editing. WB: Methodology, writing-original draft, and writing-review and editing. CL: Conceptualization, methodology, data curation, writing-review and editing, supervision, and funding acquisition. All authors read and approved the final manuscript.

Competing Interests

ARTICLE IN PRESS

All authors declare no financial or non-financial competing interests.

ARTICLE IN PRESS

ARTICLE IN PRESS

References

1 Myat, A., Song, K. J. & Rea, T. Out-of-hospital cardiac arrest: current concepts. *Lancet.* **391**, 970-979; 10.1016/S0140-6736(18)30472-0 (2018).

2 Blomberg, S. N. et al. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. *Resuscitation.* **138**, 322-329; 10.1016/j.resuscitation.2019.01.015 (2019).

3 Kurz, M. C. et al. Telecommunicator cardiopulmonary resuscitation: a policy statement from the American Heart Association. *Circulation.* **141**, e686-e700; 10.1161/CIR.0000000000000744 (2020).

4 Eberhard, K. E., Linderoth, G., Gregers, M. C. T., Lippert, F. & Folke, F. Impact of dispatcher-assisted cardiopulmonary resuscitation on neurologically intact survival in out-of-hospital cardiac arrest: a systematic review. *Scand. J. Trauma Resusc. Emerg. Med.* **29**, 70; 10.1186/s13049-021-00875-5 (2021).

5 Berg, K. M. et al. Part 7: Systems of Care: 2020 American Heart Association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. *Circulation.* **142**, S580-S604; 10.1161/CIR.0000000000000899 (2020).

6 Nguyen, D. D. et al. Association between delays in time to bystander CPR and survival for witnessed cardiac arrest in the United States. *Circ. Cardiovasc. Qual. Outcomes.* **17**, e010116; 10.1161/CIRCOUTCOMES.123.010116 (2024).

7 Viereck, S. et al. Recognising out-of-hospital cardiac arrest during emergency calls increases bystander cardiopulmonary resuscitation and survival. *Resuscitation.* **115**, 141-147; 10.1016/j.resuscitation.2017.04.006 (2017).

ARTICLE IN PRESS

8 Watkins, C. L. et al. Predictors of recognition of out of hospital cardiac arrest by emergency medical services call handlers in England: a mixed methods diagnostic accuracy study. *Scand. J. Trauma Resusc. Emerg. Med.* **29**, 7; 10.1186/s13049-020-00823-9 (2021).

9 Chen, Y. C. et al. Dispatcher-assisted cardiopulmonary resuscitation: disparity between urban and rural areas. *Emerg. Med. Int.* **2020**, 9060472; 10.1155/2020/9060472 (2020).

10 Lewis, M., Stubbs, B. A. & Eisenberg, M. S. Dispatcher-assisted cardiopulmonary resuscitation. *Circulation.* **128**, 1522-1530; 10.1161/CIRCULATIONAHA.113.002627 (2013).

11 Travers, S. et al. Out-of-hospital cardiac arrest phone detection: those who most need chest compressions are the most difficult to recognize. *Resuscitation.* **85**, 1720-1725; 10.1016/j.resuscitation.2014.09.020 (2014).

12 Juul Grabmayr, A. et al. Optimising telecommunicator recognition of out-of-hospital cardiac arrest: a scoping review. *Resusc. Plus.* **20**, 100754; 10.1016/j.resplu.2024.100754 (2024).

13 Byrsell, F. et al. Swedish dispatchers' compliance with the American Heart Association performance goals for dispatch-assisted cardiopulmonary resuscitation and its association with survival in out-of-hospital cardiac arrest: a retrospective study. *Resusc. Plus.* **9**, 100190; 10.1016/j.resplu.2021.100190 (2022).

14 Hiltunen, P. V., Silfvast, T. O., Jäntti, T. H., Kuisma, M. J. & Kurola, J. O. Emergency dispatch process and patient outcome in bystander-witnessed out-of-hospital cardiac arrest with a shockable rhythm. *Eur. J. Emerg. Med.* **22**, 266-272;

ARTICLE IN PRESS

10.1097/mej.0000000000000151 (2015).

15 Blomberg, S. N. et al. Effect of machine learning on dispatcher recognition of out-of-hospital cardiac arrest during calls to emergency medical services: a randomized clinical trial. *JAMA Netw. Open.* **4**, e2032320; 10.1001/jamanetworkopen.2020.32320 (2021).

16 Zhang, Z. & Ni, H. Critical care studies using large language models based on electronic healthcare records: a technical note. *J. Intensive Med.* **5**, 137-150; 10.1016/j.jointm.2024.09.002 (2025).

17 Hajijama, S., Juneja, D. & Nasa, P. Large language model in critical care medicine: opportunities and challenges. *Indian J. Crit. Care Med.* **28**, 523-525; 10.5005/jp-journals-10071-24743 (2024).

18 Gaber, F. et al. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *NPJ Digit. Med.* **8**, 263; 10.1038/s41746-025-01684-1 (2025).

19 Supriyono, Wibawa, A. P., Suyono & Kurniawan, F. Advancements in natural language processing: implications, challenges, and future directions. *Telemat. Inform. Rep.* **16**, 100173; 10.1016/j.teler.2024.100173 (2024).

20 Maity, S. & Saikia, M. J. Large language models in healthcare and medical applications: a review. *Bioengineering (Basel).* **12**; 10.3390/bioengineering12060631 (2025).

21 Berdowski, J., Beekhuis, F., Zwinderman, A. H., Tijssen, J. G. P. & Koster, R. W. Importance of the first link. *Circulation.* **119**, 2096-2102; 10.1161/CIRCULATIONAHA.108.768325 (2009).

22 Riou, M. et al. The linguistic and interactional factors impacting recognition and dispatch

ARTICLE IN PRESS

in emergency calls for out-of-hospital cardiac arrest: a mixed-method linguistic analysis study protocol. *BMJ Open.* **7**, e016510; 10.1136/bmjopen-2017-016510 (2017).

23 Tamminen, J. et al. Spontaneous trigger words associated with confirmed out-of-hospital cardiac arrest: a descriptive pilot study of emergency calls. *Scand. J. Trauma Resusc. Emerg. Med.* **28**, 1, doi:10.1186/s13049-019-0696-1 (2020).

24 Kirby, K., Voss, S., Bird, E. & Benger, J. Features of Emergency Medical System calls that facilitate or inhibit Emergency Medical Dispatcher recognition that a patient is in, or at imminent risk of, cardiac arrest: a systematic mixed studies review. *Resusc. Plus.* **8**, 100173; 10.1016/j.resplu.2021.100173 (2021).

25 Kirby, K., Voss, S., Benger, J. & Barnes, R. K. A conversation analytical study of call openings in Emergency Medical Service calls where the patient is at imminent risk of out-of-hospital cardiac arrest. *Resusc. Plus.* **19**, 100706; 10.1016/j.resplu.2024.100706 (2024).

26 Riou, M. et al. 'I think he's dead': a cohort study of the impact of caller declarations of death during the emergency call on bystander CPR. *Resuscitation.* **160**, 1-6; 10.1016/j.resuscitation.2021.01.001 (2021).

27 Byrsell, F. et al. Machine learning can support dispatchers to better and faster recognize out-of-hospital cardiac arrest during emergency calls: a retrospective study. *Resuscitation.* **162**, 218-226; 10.1016/j.resuscitation.2021.02.041 (2021).

28 Singhal, K. et al. Large language models encode clinical knowledge. *Nature.* **620**, 172-180; 10.1038/s41586-023-06291-2 (2023).

29 Williams, C. Y. K., Miao, B. Y., Kornblith, A. E. & Butte, A. J. Evaluating the use of

ARTICLE IN PRESS

large language models to provide clinical recommendations in the Emergency Department. *Nat. Commun.* **15**, 8236; 10.1038/s41467-024-52415-1 (2024).

30 Li, J., Monroe, W. & Jurafsky, D. Understanding neural networks through representation erasure. *arXiv.* **1**; 10.48550/arXiv.1612.08220 (2016).

31 Kohavi, R. & John, G. H. Wrappers for feature subset selection. *Artif. Intell.* **97**, 273-324; 10.1016/S0004-3702(97)00043-X (1997).

32 Collins, G. S. et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* **385**, e078378; 10.1136/bmj.q902 (2024).

33 Kim, T. H. et al. Association between patient age and pediatric cardiac arrest recognition by emergency medical dispatchers. *Am. J. Emerg. Med.* **58**, 275-280; 10.1016/j.ajem.2022.05.038 (2022).

34 Majdik, Z. P. et al. Sample size considerations for fine-tuning large language models for named entity recognition tasks: methodological study. *JMIR AI* **3**, e52095; 10.2196/52095 (2024).

35 Kalan, L., Chahine, R. A. & Lasfer, C. The effectiveness and relevance of the Canadian triage system at times of overcrowding in the emergency department of a private tertiary hospital: a United Arab Emirates (UAE) study. *Cureus.* **16**, e52921; 10.7759/cureus.52921 (2024).

36 Sparck Jones, K. A. Statistical interpretation of term specificty and its application in retrieval. *Journal of Documentation* **28**, 11-21; 10.1108/eb026526 (1972).

37 Vaswani, A. et al. Attention is all you need. *Adv Neural Inf Process Syst* **30** (2017).

ARTICLE IN PRESS

ARTICLE IN PRESS

38 Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: pre-training of deep bidirectional transformers for language understanding. *NAACL*. **1,** 4171-4186; 10.18653/v1/N19-1423 (2019).

39 Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).

40 DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837-845 (1988).

ARTICLE IN PRESS

ARTICLE IN PRESS

Figure legends

Figure 1. Performance comparison of OHCA recognition models within 60 s. **a)** AUROC values of OHCA recognition models within 60 s. **b)** AUPRC values of OHCA recognition models within 60 s. OHCA, out-of-hospital cardiac arrest; AUROC, area under the receiver-operating characteristic curve; AUPRC, area under the precision-recall curve.

Figure 2. Time to OHCA recognition across emergency calls by DyLM-OHCA. **a)** Distribution of recognition times among all predicted OHCA calls. **b)** Comparison of recognition time distributions between TP and FP predictions. The p-values were calculated based on mean comparisons between TP and FP groups. OHCA, out-of-hospital cardiac arrest; TP, true-positive; FP, false-positive.

ARTICLE IN PRESS

Figure 3. Top words contributing to OHCA recognition: Caller vs. dispatcher and prediction outcome comparison. **a)** Caller – True-positive. **b)** Dispatcher – True-positive. **c)** Caller – False-positive. **d)** Dispatcher – False-positive. OHCA, out-of-hospital cardiac arrest.

Figure 4. Sample-level word attribution for OHCA recognition using DyLM-OHCA. This figure presents the change in OHCA risk score aligned with the emergency call transcript. OHCA is detected at 10.7 s, corresponding to the caller's statement: "He's not waking up, even when I call him." The right-hand plot indicates that the risk score exceeds the detection threshold at this moment. Word-level leave-one-out analysis reveals that "not waking up" contributed most significantly to the model prediction. This phrase was produced as the English expression

ARTICLE IN PRESS

corresponding to the original Korean-language input, illustrating how the language model processes cross-lingual cues during risk estimation. OHCA, out-of-hospital cardiac arrest.

Figure 5. Flowchart of emergency calls in the study cohort.

Figure 6. DyLM-OHCA network architecture and real-time OHCA risk prediction. **a)** Timeliness module using a two-stage smoothing procedure for early OHCA detection. **b)** Real-time OHCA recognition based on dynamic risk score and first threshold crossing. OHCA, out-of-hospital cardiac arrest.

ARTICLE IN PRESS

ARTICLE IN PRESS

Figures

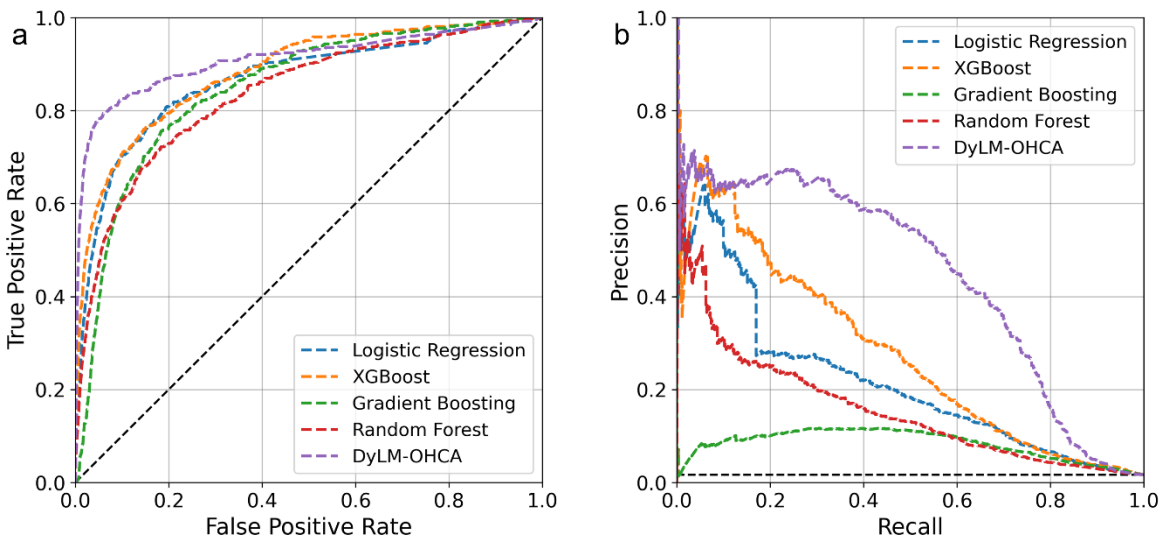


Figure 1. Performance comparison of OHCA recognition models within 60 s. **a)** AUROC values of OHCA recognition models within 60 s. **b)** AUPRC values of OHCA recognition models within 60 s.

ARTICLE IN PRESS

ARTICLE IN PRESS

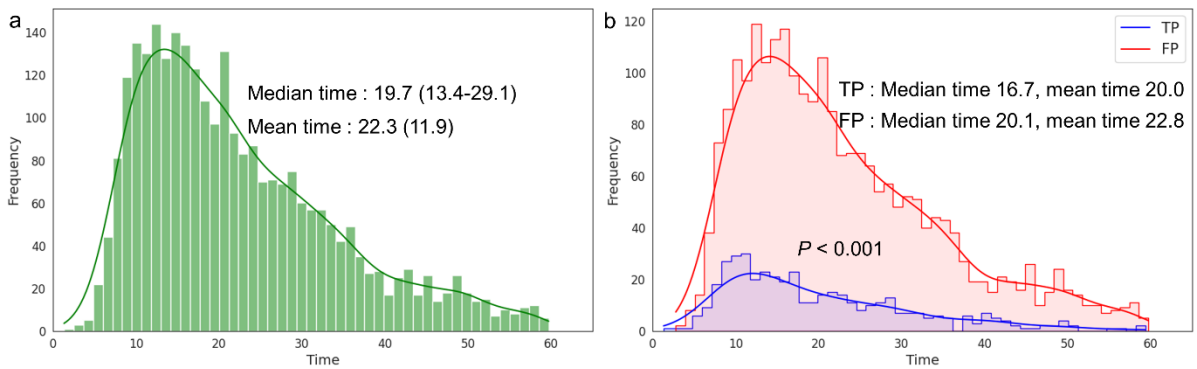


Figure 2. Time to OHCA recognition across emergency calls by DyLM-OHCA. **a)** Distribution of recognition times among all predicted OHCA calls. **b)** Comparison of recognition time distributions between TP and FP predictions. The p-values were calculated based on mean comparisons between TP and FP groups.

ARTICLE IN PRESS

ARTICLE IN PRESS

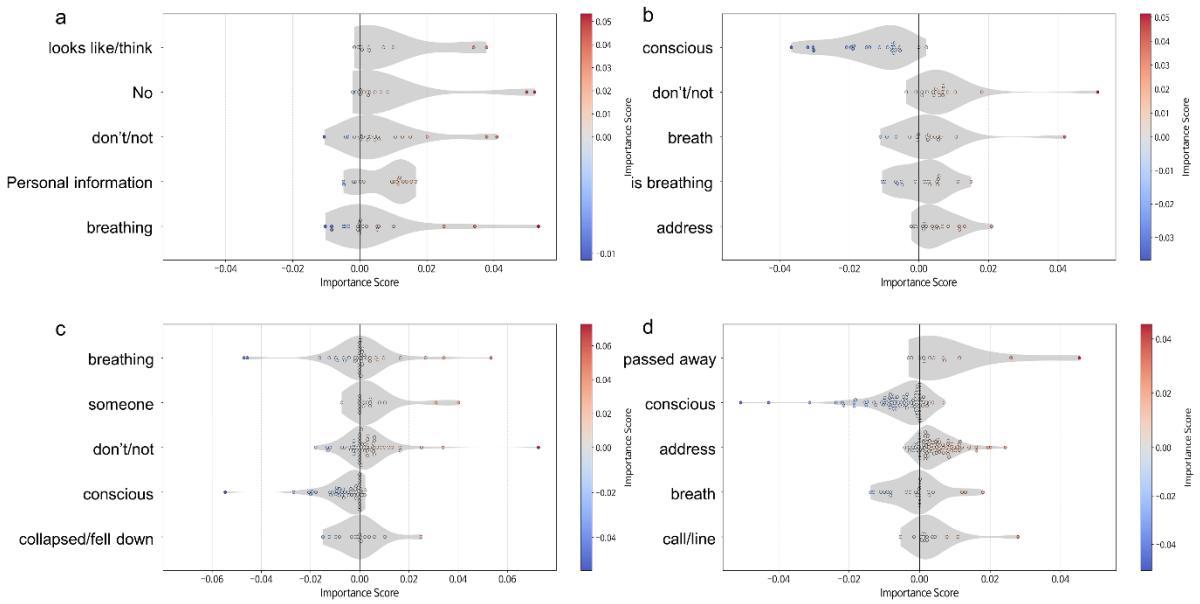


Figure 3. Top words contributing to OHCA recognition: Caller vs. dispatcher and prediction outcome comparison. **a)** Caller – True-positive. **b)** Dispatcher – True-positive. **c)** Caller – False-positive. **d)** Dispatcher – False-positive.

ARTICLE IN PRESS

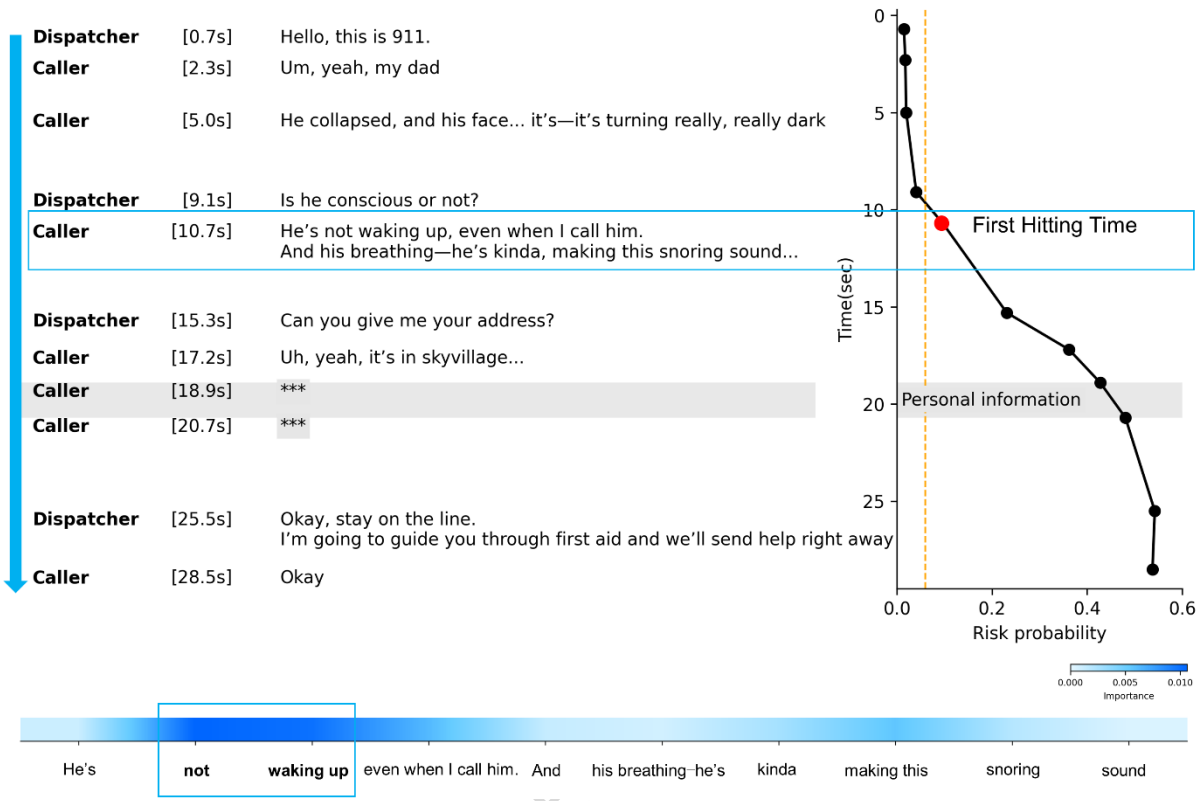


Figure 4. Sample-level word attribution for OHCA recognition using DyLM-OHCA. This figure presents the change in OHCA risk score aligned with the emergency call transcript. OHCA is detected at 10.7 s, corresponding to the caller's statement: "He's not waking up, even when I call him." The right-hand plot indicates that the risk score exceeds the detection threshold at this moment. Word-level leave-one-out analysis reveals that "not waking up" contributed most significantly to the model prediction. This phrase was produced as the English expression corresponding to the original Korean-language input, illustrating how the language model processes cross-lingual cues during risk estimation.

OHCA out-of-hospital cardiac arrest.

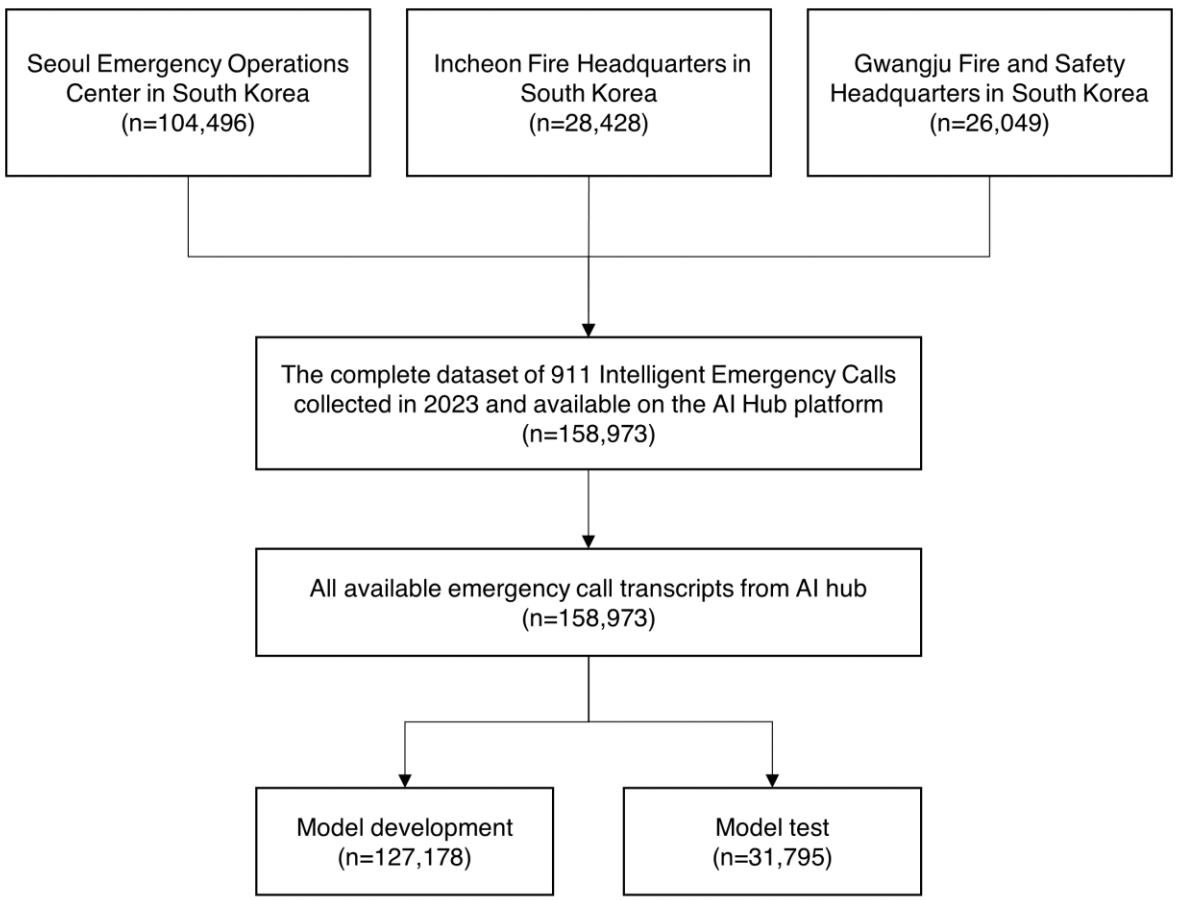


Figure 5. Flowchart of emergency calls in the study cohort.

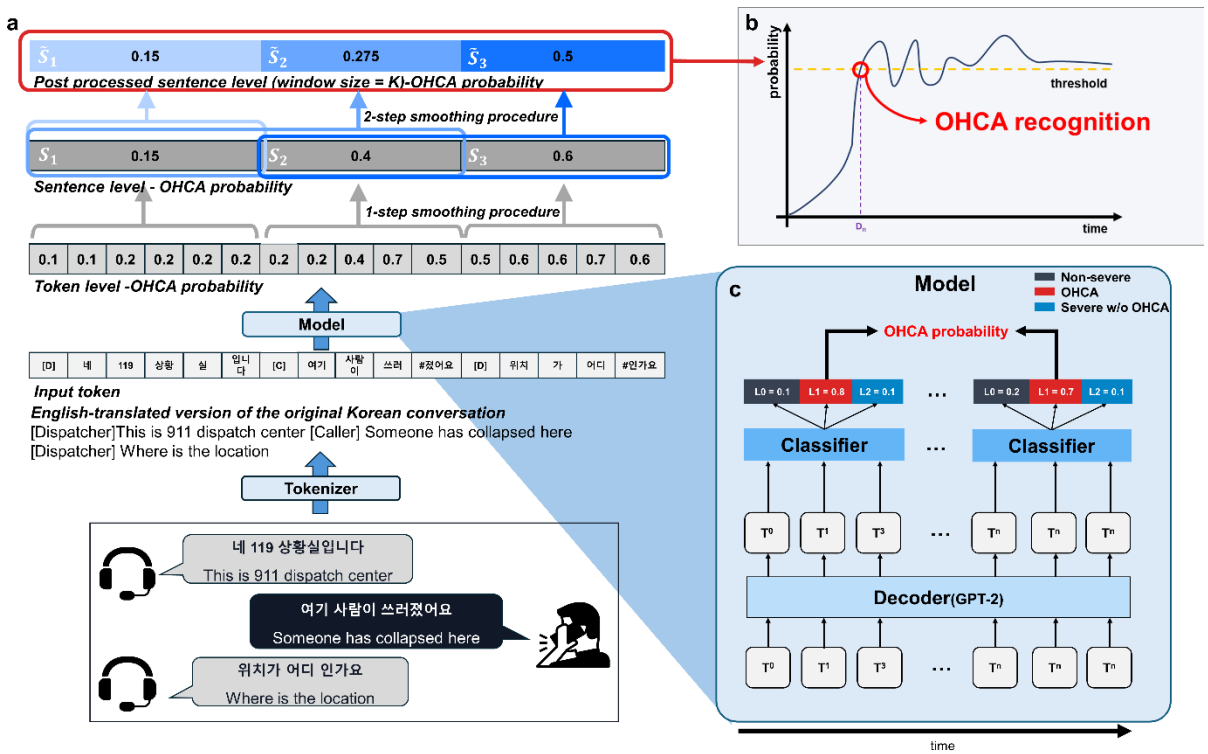


Figure 6. DyLM-OHCA network architecture and real-time OHCA risk prediction. **a)** Timeliness module using a two-stage smoothing procedure for early OHCA detection. **b)** Real-time OHCA recognition based on dynamic risk score and first threshold crossing. **c)** GPT-2 backbone (decoder-only) with causal attention to capture conversational temporal dynamics.
OHCA out-of-hospital cardiac arrest, *GPT* generative pre-trained transformer.

ARTICLE IN PRESS

Table legends

Table 1. Characteristics of emergency calls.

Table 2. Performance comparison of OHCA recognition models at recall 0.80.

Table 3. Confusion matrix of the DyLM-OHCA model (recall = 0.80).

Table 4. Performance of DyLM-OHCA across sensitivity settings: subset, caller-only, pre-protocol inputs.

Table 5. Recognition time and prediction stability in true positive vs. false positive OHCA cases.

ARTICLE IN PRESS

ARTICLE IN PRESS

Table 1. Characteristics of emergency calls.

Emergency call characteristics	Total (n = 158,973)	Train (n = 127,178)	Test (n = 31,795)	p-value
Emergency medical services				
Severe disease (CTAS 1-2)	16,528 (10.4)	13,199 (10.4)	3,329 (10.5)	0.7826
Non-severe disease (CTAS 5)	72,154 (45.4)	57,750 (45.4)	14,404 (45.3)	0.7388
Trauma	16,903 (10.6)	13,518 (10.6)	3,385 (10.6)	0.9375
Other medical cases	9,650 (6.1)	7,662 (6.0)	1,988 (6.3)	0.1312
OHCA	2646 (1.7)	2,117 (1.7)	529 (1.7)	0.6876
General accident	1,592 (1.0)	1,300 (1.0)	292 (0.9)	0.1028
Pregnancy-related cases	307 (0.2)	253 (0.2)	54 (0.2)	0.3243
Drug/chemical poisoning	259 (0.2)	208 (0.2)	51 (0.2)	0.9627
Search and rescue				
Safety accident	11,250 (7.1)	9,060 (7.1)	2,190 (6.9)	0.1455
Other rescue	7,962 (5.0)	6,363 (5.0)	1,599 (5.0)	0.8613
Property damage incident	3,201 (2.0)	2,563 (2.0)	638 (2.0)	0.9392
Suicide	404 (0.3)	329 (0.3)	75 (0.2)	0.5091
Firefighting				
General fire	9,256 (5.8)	7,400 (5.8)	1,856 (5.8)	0.9088
Other fire	1,348 (0.8)	1,091 (0.9)	257 (0.8)	0.4079
Forest fire	147 (0.1)	118 (0.1)	29 (0.1)	0.9999
Miscellaneous emergency call	5,366 (3.4)	4,247 (3.3)	1,119 (3.5)	0.1159

CTAS Canadian triage and acuity scale, *OHCA* out-of-hospital cardiac arrest.

ARTICLE IN PRESS

ARTICLE IN PRESS

Table 2. Performance comparison of OHCA recognition models at recall 0.80.

Model	ACC	AUROC	AUPRC	Recall	Precision	Specificity	F1	Brier
Logistic regression	0.808±0.002	0.865±0.006	0.213±0.010	0.801±0.012	0.065±0.003	0.808±0.002	0.120±0.005	0.015±0.000
XGBoost	0.791±0.002	0.882±0.007	0.280±0.016	0.799±0.010	0.060±0.002	0.791±0.002	0.111±0.004	0.014±0.000
Gradient Boosting	0.753±0.002	0.847±0.006	0.082±0.004	0.803±0.011	0.051±0.002	0.753±0.002	0.096±0.003	0.025±0.001
Random Forest	0.701±0.002	0.838±0.007	0.161±0.010	0.804±0.016	0.042±0.002	0.699±0.002	0.081±0.003	0.016±0.001
DyLM-OHCA	0.922±0.001	0.937±0.006	0.456±0.013	0.805±0.016	0.155±0.005	0.924±0.001	0.259±0.007	0.012±0.000

OHCA out-of-hospital cardiac arrest, *ACC* accuracy, *AUROC* area under ROC curve, *AUPRC* area under the precision-recall curve.

ARTICLE IN PRESS

ARTICLE IN PRESS

Table 3. Confusion matrix of the DyLM-OHCA model (recall = 0.80).

	Predicted OHCA	Predicted Non-OHCA
Actual OHCA	426 (TP)	103 (FN)
Actual Non-OHCA	2,345 (FP)	28,920 (TN)

OHCA out-of-hospital cardiac arrest, *TP* true positive, *FN* false negative, *FP* false positive, *TN* true negative.

ARTICLE IN PRESS

ARTICLE IN PRESS

Table 4. Performance of DyLM-OHCA across sensitivity settings: subset, caller-only, pre-protocol inputs.

Model	ACC	AUROC	AUPRC	Recall	Precision	Specificity	F1	Brier
Subset (Protocol-activated)	0.881±0.004	0.904±0.011	0.794±0.015	0.804±0.021	0.564±0.015	0.894±0.005	0.663±0.014	0.062±0.003
Subset (Excluded protocol)	0.828±0.006	0.888±0.013	0.752±0.017	0.807±0.030	0.450±0.016	0.832±0.005	0.577±0.019	0.069±0.003
Subset (Pre-protocol)	0.599±0.006	0.770±0.013	0.500±0.019	0.807±0.024	0.240±0.008	0.564±0.008	0.370±0.009	0.110±0.004
Caller-only	0.949±0.001	0.894±0.010	0.394±0.024	0.696±0.019	0.206±0.007	0.954±0.001	0.318±0.011	0.012±0.000
Pre-protocol	0.941±0.001	0.857±0.010	0.292±0.030	0.569±0.024	0.152±0.007	0.947±0.001	0.240±0.011	0.014±0.001
DyLM-OHCA	0.922±0.001	0.937±0.006	0.456±0.013	0.805±0.016	0.155±0.005	0.924±0.001	0.259±0.007	0.012±0.000

OHCA out-of-hospital cardiac arrest, *ACC* accuracy, *AUROC* area under ROC curve, *AUPRC* area under the precision-recall curve.

ARTICLE IN PRESS

ARTICLE IN PRESS

Table 5. Recognition time and prediction stability in true positive vs. false positive OHCA cases.

	True positive (n = 426)	False positive (n = 2,345)	p-value
Time to OHCA recognition (sec)	20.0 ± 11.2	22.8 ± 11.9	<0.001
Prediction stability			<0.001
Sustained cases	397 (93.2)	1,111 (47.4)	
Drop-off cases	29 (6.8)	1,234 (52.6)	

Values are mean ± standard deviation or n (%).
OHCA out-of-hospital cardiac arrest.

ARTICLE IN PRESS

ARTICLE IN PRESS

Box 1. DyLM-OHCA model-based OHCA recognition patterns.

Key dialogue trigger patterns at the first alert (FHT)

In model-positive calls (TP+FP), the first model alert commonly aligned with five recurrent dialogue patterns:

- **P1 (unresponsiveness / absent breathing)**: reports of "not breathing" or "not waking up";
- **P2 (signs suggestive of death)**: statements such as "cold" or "stiff";
- **P3 (mechanism of collapse)**: narratives involving hanging, falls, or drowning;
- **P4 (abnormal breathing descriptions)**: atypical breathing sounds (e.g., gasping or "snoring-like" breathing);
- **P5 (worsening chronic illness / end-of-life context)**: severe deterioration in patients with serious comorbidities (e.g., advanced cancer).

Operational pathways (definition and interpretation)

Operational pathways were defined as follows:

- **Caller-led**: the FHT occurred in a caller utterance, and no broad protocol screening question (consciousness or breathing) appeared before the FHT.
- **Dispatcher-led**: protocol onset preceded the FHT (i.e., at least one broad protocol screening question occurred before the FHT).
- **Workflow/context-led**: all remaining calls not meeting either definition above.

These triggers surfaced through three practical call-taking pathways:

- **Path A (caller-led)**: callers provided early, explicit descriptions consistent with P1, P2, P3, or P5, allowing an alert to occur with minimal probing.
- **Path B (dispatcher-led)**: when the initial report was vague (e.g., "collapsed" or "something is wrong"), follow-up screening elicited additional information consistent with P1 and/or P4 (and occasionally P5), after which an alert occurred.
- **Path C (workflow/context-led)**: in welfare-check or context-heavy scenarios (e.g., no contact, forced entry, "no sign of life"), the alert often coincided with rapid transitions from contextual reporting to operational steps (e.g., address confirmation) while high-risk cues (often P2, P5) were being established.

Drop-off patterns in false-positive trajectories.

In a subset of model-positive non-OHCA calls, the risk score fell below the threshold after the initial alert when subsequent utterances introduced internally inconsistent cues or pointed to alternative acute conditions (e.g., seizures or convulsions). These trajectories indicate that early high-risk cues can overlap with non-OHCA emergencies and may be revised as additional information is obtained during the call.

OHCA out-of-hospital cardiac arrest, *FHT* first hitting time, *TP* true positive, *FP* false positive