Beyond ICD Codes: Fine-Tuning LLMs for In-Hospital Cardiac Arrest Identification from EHR Notes

Davy Weissenbacher[†], PhD

Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, CA, USA E-mail: davy.weissenbacher@cshs.org

Jonathan Vo[†], MD

Center for Cardiac Arrest Prevention, Smidt Heart Institute, Cedars-Sinai Health System, Los Angeles, CA, USA E-mail: Jonathan.Vo@cshs.org

Audrey Uy-Evanado, MD

Center for Cardiac Arrest Prevention, Smidt Heart Institute, Cedars-Sinai Health System, Los Angeles, CA, USA E-mail: Audrey.Evanado@csmc.edu

Kyndaron Reinier, PhD

Center for Cardiac Arrest Prevention, Smidt Heart Institute, Cedars-Sinai Health System, Los Angeles, CA, USA E-mail: Kyndaron.Reinier@csmc.edu

Harpriya Chugh, BE, MSHS

Center for Cardiac Arrest Prevention, Smidt Heart Institute, Cedars-Sinai Health System, Los Angeles, CA, USA E-mail: Harpriya.Chugh@cshs.org

Kadiyala Vishnu, MD

Center for Cardiac Arrest Prevention, Smidt Heart Institute, Cedars-Sinai Health System, Los Angeles, CA, USA E-mail: Vishnu.Kadiyala@cshs.org

Mythreye Venkatesan, BE, MS

Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, CA, USA E-mail: Mythreye.Venkatesan@cshs.org

Karen O'Connor, MS

Perelman school of medicine, University of Pennsylvania, Philadelphia, PA, USA E-mail: karoc@pennmedicine.upenn.edu

Sumeet S. Chugh[‡], MD

Center for Cardiac Arrest Prevention, Smidt Heart Institute, Cedars-Sinai Health System, Los Angeles, CA, USA E-mail: Sumeet.Chugh@csmc.edu

Graciela Gonzalez-Hernandez[‡], PhD

Department of Computational Biomedicine, Cedars-Sinai Medical Center,

[†]Both authors contributed equally to this work

[‡]These authors jointly directed this work.

Los Angeles, CA, USA E-mail: Graciela.GonzalezHernandez@csmc.edu

Background: Identification of in-hospital cardiac arrest (IHCA) through manual chart abstraction is the gold standard, but its time-consuming and resource-intensive, which naturally limits its practicality. Diagnostic codes provide an accessible automated alternative, but prior work has shown this method of extraction has poor sensitivity and poor positive predictive value. More accurate and scalable automated methods are needed to support IHCA surveillance and research.
Methods: We developed large language model (LLM)-based systems that incorporated physician-defined clinical knowledge through guideline-enriched prompting, supervised fine-tuning, and a correction procedure. The systems identified IHCA events and their locations from electronic health record notes. Performance was evaluated at the encounter level against physician-adjudicated chart review and compared with ICD-code–based identification.
Results: Compared with an ICD-code–based algorithm (precision 0.68, recall 0.91, F1 score 0.78), the supervised fine-tuned LLM offered substantially better performance for IHCA identification on a curated evaluation corpus, reaching a precision of 0.88, a recall of 1.00, and an F1 score of 0.94. When tested on a larger, unselected validation cohort of 45,525 encounters, a separate LLM configuration—relying on guideline-enriched prompting and a self-correction procedure rather than fine-tuning—achieved a precision of 0.79 for encounter-level IHCA identification, meaning most encounters flagged by the model were genuine IHCA cases. The fine-tuned LLM underperformed on this cohort most likely due to its inability to self-correct its decisions. Using a separate model for the correction step, we mitigated this issue and reached a high precision of 0.94.
Conclusion: Explicit integration of clinical knowledge through prompting and supervised fine-tuning enabled accurate identification of IHCA events and locations from routine clinical documentation. In a two-pass workflow, these systems could prioritize encounters for clinician adjudication, reduce manual screening requirements, and support more scalable IHCA surveillance and registry development.

Keywords: In-Hospital Cardiac Arrest, Sudden Cardiac Arrest, Information Extraction, Large Language Models, Generative Systems, Supervise Fine-Tuning

1. Introduction

In-hospital cardiac arrest (IHCA) affects approximately 300,000 patients annually in the United States, corresponding to 1.6 to 2.8 events per 1000 hospital admissions.[1] Survival to hospital discharge remains no higher than approximately 25%, underscoring the substantial mortality associated with these events.[2] IHCA is conventionally defined as cessation of cardiac activity in a hospitalized patient requiring chest compressions and/or defibrillation as part of cardiopulmonary resuscitation (CPR).[3]

Despite advances in resuscitation care, the incidence of IHCA has continued to increase, while improvements in survival have recently plateaued.[1,4] These trends highlight the need

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

for accurate surveillance systems that can support benchmarking, outcomes research, and data-driven quality improvement. Accordingly, the American Heart Association's "Ten Steps Toward Improving In-Hospital Cardiac Arrest Quality of Care and Outcomes" identifies the development of high-quality IHCA registries as a central component of efforts to improve resuscitation care.[5]

Current IHCA research relies heavily on prospective registries in which trained personnel manually abstract information from medical records and cardiac arrest flow sheets. The largest such registry in the United States is the American Heart Association's *Get With The Guidelines–Resuscitation* registry.[6] However, only 5–6% of US hospitals participate in the registry in a given year, limiting the representativeness and reach of registry-based surveillance.[7] Expanding participation will require methods that preserve the clinical validity of expert adjudication while reducing the associated abstraction burden.

Outside prospective registries, IHCA cohorts are generally identified through manual chart review or through International Classification of Diseases (ICD) diagnosis and procedure codes extracted from the electronic health record (EHR).[8,9] Manual review provides clinically detailed case ascertainment but requires substantial time and specialized expertise. ICD-based methods are considerably more scalable, but they have demonstrated limited sensitivity and positive predictive value for IHCA because of missing codes, miscoding, and codes carried forward from previous encounters.[8,9] Regional and temporal variation in coding practices further limits their reliability for longitudinal surveillance and interinstitutional benchmarking.[10]

Clinical notes provide a richer representation of the events, interventions, and contextual information needed to determine whether an IHCA occurred. Natural language processing (NLP) and large language models (LLMs) have been applied to clinical text to identify cardiovascular phenotypes and risk factors, including heart failure, recurrent atrial fibrillation, and heart disease risk factors.[11–14] To our knowledge, however, prior work has not addressed automated IHCA case identification from unfiltered clinical documentation. This task requires more than recognition of cardiac-arrest terminology: a system must apply clinically defined criteria, distinguish current events from historical or hypothetical mentions, determine whether resuscitative treatment occurred, and resolve whether the event began in an inpatient unit, the emergency department, or outside the hospital.

In this study, we investigate two complementary approaches for integrating physician-defined IHCA knowledge into LLMs. First, we explicitly encode clinical definitions and annotation rules through guideline-enriched prompting and few-shot examples. Second, we incorporate the same clinically adjudicated knowledge into the model parameters through supervised fine-tuning and evaluate and refine its predictions using external-model or self-correction procedures. Although these methods are widely used in this rapidly evolving field, our contribution lies in their systematic combination and evaluation for IHCA detection, particularly across curated and real-world cohorts.[15] We develop and evaluate LLM-based systems for identifying IHCA events and their locations from unfiltered EHR notes, using physician-adjudicated chart abstraction as the reference standard and ICD-code-based identification as the principal comparator.

2. Materials and Methods

2.1. *Data Compilation and Annotation*

This study was reviewed and approved by the Cedars-Sinai Medical Center Institutional Review Board. We constructed two datasets for model development and evaluation. The first, termed the *Evaluation Corpus*, comprised 500 clinical encounters, defined as a complete episode of care from intake through discharge, from the Cedars-Sinai Medical Center electronic health record (EHR). An individual patient could contribute more than one encounter. For each encounter, the corpus included the available inpatient documentation, including emergency department (ED) notes, discharge summaries, and Code Blue notes recorded following cardiac arrest resuscitation. An encounter could contain multiple notes of the same type.

To obtain sufficient positive examples for model development, the Evaluation Corpus combined code-enriched and randomly sampled encounters. Of the 500 encounters, 47% contained at least one diagnosis or procedure code selected to identify a potential cardiac arrest, while the remaining 53% were randomly sampled encounters. Code presence was used only for cohort enrichment and was not treated as evidence that an IHCA had occurred. The codes used for sampling are listed in Table 1.

Table 1: Diagnosis and procedure codes used to enrich the Evaluation Corpus for potential cardiac arrest encounters

Code	Description
I49.01	Ventricular fibrillation
I49.02	Ventricular flutter
I46.2	Cardiac arrest due to underlying cardiac condition
I46.8	Cardiac arrest due to other underlying condition
I46.9	Cardiac arrest, cause unspecified
92950	Cardiopulmonary resuscitation
5A12012	Performance of cardiac output, single, manual
02QA0ZZ	Repair heart, open approach

Before annotation began, the two lead investigators iteratively developed annotation guidelines grounded in the Utstein definition of IHCA as pulselessness requiring chest compressions and/or defibrillation in a hospitalized patient.[3] Two physicians (JV and AUE) independently applied the draft guidelines to a pilot sample of 50 notes. They adjudicated disagreements and refined the guidelines until no further modifications were required. The final annotation guidelines were previously published by Vo et al.[16]

Using these guidelines, JV and AUE annotated 3,822 notes from the 500 encounters. The physician annotations served as the reference standard. A note was labeled *IHCA* when the clinical text clearly documented a pulseless event treated with chest compressions and/or external defibrillation; otherwise, it was labeled *not IHCA*. Labels were based on whether resuscitative treatment was delivered rather than on whether treatment was clinically indicated or concordant with the patient's code status. Thus, a pulseless event was labeled *IHCA* when

chest compressions and/or external defibrillation were performed, including when the patient had a pre-existing do-not-resuscitate or comfort-care order. A pulseless event for which resuscitation was not attempted was labeled *not IHCA*. A note was classified as positive for IHCA only when it documented both the occurrence of cardiac arrest or pulselessness and the delivery of resuscitative treatment, such as chest compressions or defibrillation. In the rare cases in which these two criteria were documented separately across multiple notes, the presence of multiple notes for the same patient generally ensured that at least one note documented both criteria together, thereby minimizing the risk of missing an IHCA.

Identifying all episodes of in-hospital pulselessness, including events without attempted resuscitation, represents an important but distinct surveillance objective relevant to end-of-life care and adherence to treatment limitations. Such events were outside the scope of this study, which was designed to align with Utstein-style IHCA registries focused on cardiac arrests for which resuscitation was attempted.[3]

For each note labeled positive for IHCA, we also annotated the location of the arrest. Location was categorized as: 1) in-hospital (IH), including critical care units, operating rooms, and cardiac catheterization laboratories; 2) emergency department (ED); and 3) out-of-hospital cardiac arrest for which treatment was initiated out of the hospital and continued on arrival to the emergency department (OH-ED). The third category, OH-ED (out-of-hospital cardiac arrest with continued treatment in the ED), is not conventionally included in standard IHCA definitions; its inclusion here was intended to facilitate identification of this subgroup and permit its exclusion in future research, as appropriate. For notes documenting multiple arrests, only the first arrest was labeled.

Following annotation of each individual note within an encounter, two physicians reviewed all notes for each encounter to adjudicate IHCA status and location at the encounter level. If there were multiple IHCA events in one encounter, the encounter-level location was adjudicated from the first event.

We randomly split the Evaluation Corpus into a training set (300 encounters, 60%), a development set (50 encounters, 10%), and a test set (150 encounters, 30%). The training set included 99 encounters with at least one IHCA event, comprising 172 notes labeled positive for IHCA. Of these 99 encounters, 6% were further classified as OH-ED (6 encounters, 12 notes), 17% as ED (17 encounters, 37 notes), and 77% as IH (76 encounters, 123 notes). The development set included 11 positive encounters, comprising 22 IHCA-positive notes. Of these encounters, 18% were classified as OH-ED (2 encounters, 5 notes), 9% as ED (1 encounter, 3 notes), and 73% as IH (8 encounters, 14 notes). Finally, the test set included 37 positive encounters, comprising 70 IHCA-positive notes: 8% of these encounters were classified as OH-ED (3 encounters, 6 notes), 16% as ED (6 encounters, 16 notes), and 76% as IH (28 encounters, 48 notes). Two physicians independently annotated the same 50 encounters for presence and location of IHCA, noting a Cohen's kappa of 0.951 for IHCA presence (almost perfect agreement) and 0.838 for IHCA location (strong agreement).[17]

Because the Evaluation Corpus was enriched with potential IHCA cases identified using diagnosis and procedure codes, its IHCA prevalence was higher than that expected in routine hospital data. We therefore conducted an additional evaluation using a year of encounters

from the ongoing Observational Study of Sudden Cardiac Arrest (OSCAR) cohort.[18] (dated between January 1 and December 31, 2023 (n = 45,525)), including inpatient, emergency department, and observation encounters, as well as encounters containing a Code Blue note. Ambulatory encounters were excluded. No additional code-based enrichment or case preselection was applied.

The LLM was applied to all encounters in the 2023 OSCAR subset, and encounters predicted as IHCA were reviewed by two physicians (JV and AUE).[18] For each model-flagged encounter, the physicians reviewed all flagged notes and determined whether the event met the Utstein definition of IHCA.[3,18] Encounters meeting this definition were classified as true positives, whereas the remaining model-flagged encounters were classified as false positives. Because model-negative encounters were not systematically adjudicated, this analysis was designed to estimate precision rather than recall or overall classification accuracy.

2.2. *LLM choice and prompt development*

As an initial step, we developed a strong off-the-shelf, prompt-based system to establish a performance baseline before supervised fine-tuning. We selected Meta's LLaMA 3.1-405B instruction-tuned model because it is an open-weight model available with strong general instruction-following capabilities that can be deployed within on-premises computing environment, although it requires substantial computational resources.[19]

We evaluated different approaches, beginning with two prompt-based baselines. In the first baseline approach (line 2 in Table 2), the LLM was given a role description, a concise statement of the task (classifying notes mentioning IHCA and location), and the note text. In the second baseline (line 3 in Table 2), we augmented this prompt with a summary of the clinical guidelines used for annotation, concise definitions of clinical concepts, and rule-based restatements of the guidelines to compensate for the model's limited exposure to clinician-authored medical documents. We then enhanced this second baseline using few-shot and chain-of-thought prompting.[20] Few-shot prompting was implemented by adding a 10 training examples with correct IHCA and location labels, selected through iterative error analysis to highlight common false positives as well as representative true positives and true negatives. Chain-of-thought prompting, asking the LLM to briefly justify its predictions and cite supporting phrases, aka quotes, from the note, promotes step-by-step reasoning to improve explainability and encourage grounding.[20] Finally (line 4 in Table 2) added a self-correction step on top of this enhanced prompt. For each note, we first obtained the model's classification and reasoning, then triggered a second review only for notes initially classified as IHCA, as most remaining errors involved false positives or misassigned locations. In this review, the model received the initial classification, location, reasoning, and note text, and was instructed to confirm or revise its decision using one of three location-specific prompts (IH, ED, or OH-ED).

2.3. *Supervised full fine-tuning LLM*

In the second step of our approach, we investigated training the model to further improve performance. We deliberately selected a small model so that full-parameter supervised fine-tuning could be conducted within our available computational resources. This allowed us

to study direct task adaptation by updating all model weights, while maintaining a training procedure that was computationally feasible and reproducible. We did not evaluate parameter-efficient fine-tuning methods such as LoRA or preference-optimization approaches such as RLHF, as these would have required additional method-specific hyperparameter tuning and experiments beyond the scope of this study. Such approaches may be more appropriate for larger models and remain important directions for future work.

With numerous open-source LLMs now available spanning different sizes and performance levels, we chose the Qwen 3 series[21] for our experiments. This series includes both dense and mixture-of-experts models ranging from 0.6 to 235 billion parameters and is specifically optimized for reasoning. Computational and time constraints led us to select the smaller Qwen3-4B Instruct variant. We conducted fine-tuning using the Transformer Reinforcement Learning (TRL) library,[22] chosen for its integration with Hugging Face Transformers.

To train the model, we presented each note and, using the same final prompt as for the system in line 3 of Table 2, asked the model to generate a JSON object indicating whether the note mentioned an in-hospital cardiac arrest and, if so, where it occurred (IH, ED, or OH-ED), along with supporting quotes from the text and a short explanation for its decision. We corrected the model's output whenever it diverged from labels assigned by our physicians; for notes labeled as mentioning an in-hospital cardiac arrest, we provided a single standardized explanation—"The patient experienced a cardiac arrest and received treatment for it"—along with the supporting quotes as the correction. We fine-tuned the model for eight epochs on four NVIDIA A100 GPUs using a learning rate of 1×10^{-5} and an effective batch size of 16 examples per optimization step. The effective batch size was obtained using a per-device batch size of 2 and two gradient-accumulation steps. Training used BF16 mixed precision. We used TRL's default token-level cross-entropy loss, leaving all other training parameters at their default values. We selected the final checkpoint produced at the end of training.

2.4. *Testing and validation*

We evaluated the performance of each baseline and the final system on the test set of the Evaluation Corpus at the encounter level. Encounter-level IHCA predictions were chosen over note-level predictions because clinical and research applications depend on whether a patient experienced IHCA during an encounter, rather than whether an individual note mentions IHCA, and because ICD codes also identify IHCA at the encounter level.

As a conventional NLP baseline, we compared the performance of our trained model against a simple baseline (line 5): the out-of-the-box Qwen3-4B Instruct model, prompted with our final prompt and our 10 few-shot examples. We also fine-tuned Bio_ClinicalBERT,[23] a BERT-based encoder pretrained on biomedical literature and MIMIC-III clinical notes, as a four-class note-level classifier. Because its 512-token context window was exceeded by a 20% of the notes, longer notes were divided into overlapping 512-token windows with a 128-token stride and window level predictions were aggregated to the note level. Note-level predictions were subsequently aggregated at the encounter level using the same procedure as for the other systems. The model was fine-tuned for eight epochs with an effective batch size of 16, a learning rate of 2×10^{-5}, BF16 precision, and inverse-square-root-frequency class weighting,

with the checkpoint selected by encounter-level macro F1 on the development set.

We compared models performance to performance of ICD codes commonly used to identify IHCA (ICD codes I49.01, I49.02, I46.2, I46.8, I46.9, CPT code 92950, and procedure codes 5A12012 and 02QA0ZZ). ICD performance was evaluated in two steps. To assess recall, these codes were applied to a gold-standard list of known IHCA encounters from the annotated corpus, and any positive encounters not captured by the codes were counted as false negatives. To assess precision, we applied the same codes to encounters from the larger OSCAR study between 7/1/2021 and 10/31/2021 (2021 OSCAR subset), and a physician adjudicated each candidate as IHCA or not; encounters adjudicated as not IHCA were counted as false positives.

We summarized binary encounter-level performance for IHCA detection using precision, recall, and F1 score: precision (equivalent to positive predictive value) denotes the proportion of predicted IHCAs that are true IHCAs, recall (equivalent to sensitivity) denotes the proportion of true IHCAs correctly identified, and F1 score denotes the harmonic mean of the two. Multiclass performance for IHCA location was assessed using a confusion matrix, with correct predictions appearing on the diagonal (see Figure 1).[24]

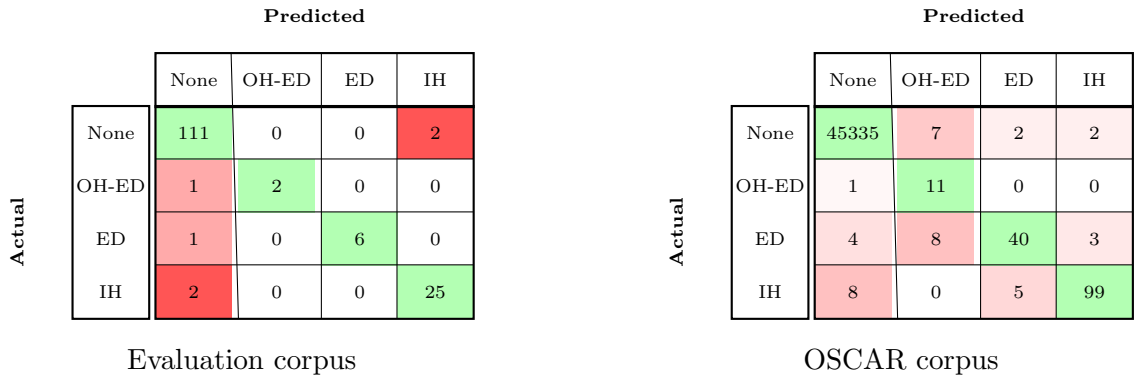
3. Results

Table 2 details the performance of all systems. First, all generative systems achieve higher F1 scores than the ICD-based system (line 1) and Bio_ClinicalBERT (line 8) on the test set of the Evaluation corpus, regardless of the settings used to generate answers. The Bio_ClinicalBERT baseline (line 8) achieved an encounter-level precision of 0.62, recall of 1.00, and F1 score of 0.76 on the Evaluation corpus test set. In preliminary experiments, inverse-frequency class weighting yielded the same recall of 1.00 but substantially lower precision (0.29) and F1 score (0.45); inverse-square-root-frequency weighting was therefore used for the final Bio_ClinicalBERT baseline.

Second, prompt refinement produced a six-point gain in F1 for Llama 405B with few-shot prompting (line 3), compared with the baseline prompt (line 2), on the Evaluation corpus test set. Adding the self-correction step then reduced recall from 1.0 to 0.97 (line 3 vs. line 4), while precision increased from 0.86 to 0.88 on the Evaluation corpus and from 0.47 to 0.79 on the OSCAR corpus. Finally, full fine-tuning supervision produced the largest F1 gains: the best-performing system (line 6) achieved perfect recall on the Evaluation corpus test set and an F1 score of 0.94, compared to 0.92 for LLaMA-405B, a model 100 times larger. On the OSCAR corpus, precision for this system achieve a low 0.59 compared to 0.79 for LLaMA-405B. Note that, on the Evaluation corpus, although the fine-tuned Qwen3-4B model (line 6) outperformed LLaMA-405B (line 4), the difference did not reach statistical significance (paired bootstrap on binary note-level classification: 95% CI [-0.002, 0.097], $p = 0.068$). Despite a substantial overall sample size (N=1203), the relatively small number of positive cases (n=70) limited statistical power to detect differences in F1 score between systems. By contrast, on the OSCAR corpus, the LLaMA-405B (line 4) achieved significantly higher precision than the fine-tuned Qwen3-4B model (line 6): (paired bootstrap on binary note-level classification, restricted to notes flagged by at least one system: 95% CI [-0.142, -0.067], $p < 0.001$).

Table 2: IHCA detection performance at the Encounter-level: Evaluation and OSCAR 2023 Corpora. FS stands for few-shots, SFT for supervised fine-tuning, SC for self-correction, LC for Llama correction.

	Evaluation corpus			OSCAR 2023 corpus
	Prec.	Rec.	F1	Prec.
1. ICD Codes	0.68	0.91	0.78	—
2. Llama 405B - Basic Prompt	0.77	**1.0**	0.87	—
3. Llama 405B - Final Prompt + FS	0.86	**1.0**	0.93	0.47
4. Llama 405B - Final Prompt + FS + SC	0.88	0.97	0.92	0.79
5. Qwen 3 4B - Final Prompt + FS + SC	0.86	0.73	0.79	—
6. Qwen 3 4B - Final Prompt + SFT + SC	0.88	**1.00**	**0.94**	0.59
7. Qwen 3 4B - Final Prompt + SFT + LC	**0.94**	0.89	0.92	**0.94**
8. Bio_ClinicalBERT	0.62	**1.0**	0.76	—



Evaluation corpus OSCAR corpus

Fig. 1: Confusion matrices for encounter-level IHCA prediction by the fine-tuned Qwen3-4B model with Llama correction (line 7 in Table 2) on the test set of the Evaluation Corpus (left) and the OSCAR Corpus (right). Values along the diagonal indicate correct predictions; values in red indicate incorrect predictions. None refers to no IHCA. OH-ED refers to SCA that begins outside of the hospital with treatment continuing in the ED. ED refers to arrest that begins in the ED. IH refers to arrest that begins while admitted to the hospital.

4. Discussion

This study demonstrates that physician-defined clinical knowledge can be incorporated into LLM-based IHCA identification through both contextual and parametric adaptation. Guideline-enriched prompting, few-shot examples, and targeted self-correction substantially improved the performance of an off-the-shelf LLaMA 3.1-405B model, while supervised full fine-tuning enabled a much smaller Qwen3-4B model to achieve the best overall performance. On the Evaluation Corpus, the fine-tuned Qwen3-4B model achieved a precision of 0.88, recall of 1.00, and F1 score of 0.94. These findings extend our previous work on LLM-based IHCA identification by showing that task-specific supervised adaptation can encode clinically

adjudicated definitions and decision rules within a smaller model.[16]

The progression across prompt configurations also illustrates the value of explicitly representing clinical knowledge. The basic prompt required the model to classify a note using only a role description and task statement. Adding the physician-developed annotation guidelines, clinical definitions, selected examples, and evidence-grounded rationale instructions increased F1 from 0.87 to 0.93. The self-correction step produced a modest reduction in recall but improved precision, particularly in the lower-prevalence OSCAR cohort. These results suggest that explicit decision criteria and targeted review of likely error cases can improve discrimination between true IHCA events and clinically similar documentation, including historical arrest mentions, ED arrests, and out-of-hospital arrests with continued treatment in the ED.

Supervised fine-tuning produced an additional improvement over prompting alone. The fine-tuned Qwen3-4B model achieved an F1 score of 0.94, compared with 0.92 for the prompted LLaMA 3.1-405B model with self-correction. Although this difference did not reach statistical significance on the Evaluation Corpus, the finding is operationally relevant: a 4-billion-parameter model may be more feasible for secure local deployment than a 405-billion-parameter model. The result also indicates that task-specific clinical supervision can partially compensate for differences in model scale.

The OSCAR analysis addressed a different question from the curated Evaluation Corpus. Rather than estimating complete classification performance, it evaluated the precision of model-positive predictions in a year-long cohort without code-based enrichment. The prompted LLaMA model's precision increased from 0.47 to 0.79 after self-correction, indicating that the second-pass review substantially reduced false-positive screening results. The final precision of the fine-tuned Qwen3-4B model was lower with 0.59. Because model-negative OSCAR encounters were not systematically adjudicated, recall, overall accuracy, and the underlying IHCA prevalence cannot be estimated from this analysis. Accordingly, the OSCAR findings support the model's potential utility for prioritizing encounters for review but do not constitute external validation or a complete population-level performance assessment.

These findings have practical implications for IHCA surveillance. Manual abstraction remains necessary for high-validity registry construction but requires substantial clinical effort.[2] A two-stage workflow could use the model to screen routine documentation and present likely IHCA encounters, together with supporting text, for clinician adjudication. Such an approach could reduce the number of records requiring full review while preserving human oversight. However, the present study did not directly measure abstraction time, number needed to review, reviewer agreement with model-provided evidence, computational cost, or workflow integration. These operational outcomes should be evaluated prospectively before claims of reduced workload or improved registry participation can be confirmed. If validated across institutions, such systems could lower barriers to participation in IHCA surveillance and quality-improvement initiatives.[5]

The study also highlights limitations in the clinical knowledge available to general-purpose LLMs. Publicly available models are trained predominantly on nonclinical text, and their performance may be constrained by limited exposure to clinical documentation and institution-specific language.[25,26] Guideline-enriched prompting partially addresses this limitation by sup-

plying relevant knowledge at inference time, while supervised fine-tuning incorporates task-specific knowledge into model parameters. Further adaptation using clinical documents and reinforcement learning with human-feedback may improve robustness across note types and documentation styles.[27,28]

Several limitations warrant emphasis. First, all data were obtained from a single health system. Documentation conventions, Code Blue workflows, note templates, and terminology may differ across institutions and EHR platforms. Evaluation at independent health systems is therefore necessary before generalizability can be established.

Second, the Evaluation Corpus was intentionally enriched with encounters containing cardiac-arrest-related codes and included relatively few positive encounters in the test set. Although this corpus allowed us to estimate recall because all encounters were adjudicated, its sampling design and limited number of positive cases introduce uncertainty; therefore, the observed recall of 1.00 should be interpreted cautiously and may not generalize to the OSCAR population. Adjudicating a sufficiently large sample of model-negative encounters in OSCAR would be an important direction for future work.

Third, only model-positive encounters were adjudicated in the 2023 OSCAR analysis, limiting that evaluation to precision.

Fourth, the training targets used a standardized explanation rather than physician-authored reasoning tailored to each example. Consequently, the generated rationales and supporting quotations were not evaluated for completeness, faithfulness, or clinical usefulness and should not be interpreted as validated explanations. Richer supervision or human-feedback methods may improve evidence selection, but these approaches would require separate evaluation.

Fifth, the relatively poor performance of the fine-tuned Qwen3-4B with self-correction on the OSCAR corpus (line 4) was initially surprising. Our analysis revealed that the fine-tuned Qwen3-4B was largely unable to auto-correct its answers: it almost always confirmed its original decision, and in the rare cases where it did revise its output, the change was limited to the location of the cardiac arrest rather than the underlying classification. This resulted in a large number of false positives that persisted despite the self-correction step—errors that Llama-405B, by contrast, was able to correctly resolve during its own self-correction. We hypothesize that this reflects a loss of calibration following task-specific fine-tuning: as the model became increasingly specialized in detecting IHCA within the training distribution, it also grew overconfident in its predictions, which is consistent with catastrophic forgetting of the general self-critique behavior present in the base instruction-tuned model as supported by our series of experiments in.[28] To address this, we tested an alternative pipeline in which Llama-405B performs the correction step on the fine-tuned model's outputs, rather than the fine-tuned model correcting itself. This pipeline improved overall performance on the OSCAR corpus, achieving the best precision of 0.94. On the Evaluation Corpus, it likewise improved precision to 0.94 without substantially reducing recall (0.89), maintaining comparable overall performance to the fine-tuned Qwen3-4B with self-correction, with an F1 score of 0.92.

Finally, the present results reflect retrospective model development. Prospective implementation would require secure data handling, access controls, audit trails, monitoring for

performance drift, and clearly defined clinician responsibility. The model should be used to support, rather than replace, physician adjudication.

5. Conclusion

Explicit integration of physician-defined clinical knowledge through guideline-enriched prompting, supervised fine-tuning, and targeted correction enabled accurate identification of IHCA events and their locations from EHR notes. A fully fine-tuned 4-billion-parameter model achieved the strongest performance to a substantially larger prompted model.

These findings support the use of locally deployable LLMs as screening tools for retrospective IHCA cohort construction. In a clinician-supervised workflow, the model could prioritize likely IHCA encounters for adjudication and reduce reliance on diagnosis and procedure codes alone. Validation across independent health systems, complete assessment in lower-prevalence cohorts, evaluation of model-provided evidence, and prospective measurement of abstraction efficiency are required before operational deployment or broad registry use.

Data Availability

Prompts are available upon reasonable request.

Code Availability

Code available upon reasonable request.

Competing Interests

The authors declare no competing interests.

Funding

This work was supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM014731.

References

1. L. Wu, B. Narasimhan, K. Bhatia *et al.*, Temporal trends in characteristics and outcomes associated with in-hospital cardiac arrest: A 20-year analysis (1999-2018), *J Am Heart Assoc* **10**, p. e021572 (2021).
2. M. J. Holmberg, C. E. Ross, G. M. Fitzmaurice, P. S. Chan, J. Duval-Arnould, A. V. Grossestreuer, T. Yankama, M. W. Donnino and L. W. Andersen, Annual incidence of adult and pediatric in-hospital cardiac arrest in the united states, *Circ Cardiovasc Qual Outcomes* **12**, p. e005580 (2019).
3. J. P. Nolan, R. A. Berg, L. W. Andersen, F. Bhanji, P. S. Chan, M. W. Donnino, S. H. Lim, M. H. Ma, V. M. Nadkarni, M. A. Starks, G. D. Perkins, P. T. Morley and J. Soar, Cardiac arrest and cardiopulmonary resuscitation outcome reports: Update of the utstein resuscitation registry template for in-hospital cardiac arrest: A consensus report from a task force of the international

liaison committee on resuscitation (american heart association, european resuscitation council, australian and new zealand council on resuscitation, heart and stroke foundation of canada, interamerican heart foundation, resuscitation council of southern africa, resuscitation council of asia), *Resuscitation* **144**, 166 (2019).
4. J. Vo, F. L. Norby, P. Marano, Y. Matusov, K. Reinier, J. Ebinger, H. Halperin and S. S. Chugh, Management and prevention of in-hospital cardiac arrest: present and future, *npj Cardiovascular Health* **1**, p. 7 (2024).
5. B. K. Nallamothu, R. Greif, T. Anderson, H. Atiq, T. B. Couto, J. Considine, A. R. De Caen, T. Djärv, A. Doll, M. J. Douma, D. P. Edelson, F. Xu, J. C. Finn, G. Firestone, S. Girotra, K. G. Lauridsen, C. K. Leong, S. H. Lim, P. T. Morley, L. J. Morrison, A. Moskowitz, A. Mullasari Sankardas, M. T. M. Mohamed, M. C. Myburgh, V. M. Nadkarni, R. W. Neumar, J. P. Nolan, J. Athieno Odakha, T. M. Olasveengen, J. Orosz, G. D. Perkins, J. K. Previdi, C. Vaillancourt, W. H. Montgomery, C. Sasson and P. S. Chan, Ten steps toward improving in-hospital cardiac arrest quality of care and outcomes, *Circ Cardiovasc Qual Outcomes* **16**, p. e010491 (2023).
6. M. A. Peberdy, W. Kaye, J. P. Ornato, G. L. Larkin, V. Nadkarni, M. E. Mancini, R. A. Berg, G. Nichol and T. Lane-Trultt, Cardiopulmonary resuscitation of adults in the hospital: A report of 14720 cardiac arrests from the national registry of cardiopulmonary resuscitation, *Resuscitation* **58**, 297 (2003).
7. A. Uzendu, S. Girotra and P. S. Chan, What cardiac arrest registries can tell us about health differences and disparities, *Resusc Plus* **18**, p. 100614 (2024).
8. C. DeZorzi, B. Boyle, A. Qazi, K. Luthra, R. Khera, P. S. Chan and S. Girotra, Administrative billing codes for identifying patients with cardiac arrest, *J Am Coll Cardiol* **73**, 1598 (2019).
9. R. Khera, J. A. Spertus, M. A. Starks, Y. Tang, S. M. Bradley, S. Girotra and P. S. Chan, Administrative codes for capturing in-hospital cardiac arrest, *JAMA Cardiol* **2**, 1275 (2017).
10. S. J. Nelson, Y. Yin, E. A. Trujillo Rivera, Y. Shao, P. Ma, M. S. Tuttle, J. Garvin and Q. Zeng-Treitler, Are icd codes reliable for observational studies? assessing coding consistency for data quality, *Digit Health* **10**, p. 20552076241297056 (2024).
11. J. W. Cunningham, P. Singh, C. Reeder, E. S. Lau, S. Khurshid, X. Wang, P. T. Ellinor, S. A. Lubitz, P. Batra and J. E. Ho, Natural language processing for adjudication of heart failure in the electronic health record, *JACC Heart Fail* **11**, 852 (2023).
12. E. H. Houssein, R. E. Mohamed and A. A. Ali, Heart disease risk factors detection from electronic health records using advanced nlp and deep learning techniques, *Scientific Reports* **13**, p. 7173 (2023).
13. P. M. Marti-Castellote, C. Reeder, B. Claggett, P. Singh, E. S. Lau, S. Khurshid, P. Batra, S. A. Lubitz, M. Maddah, O. Vardeny, E. F. Lewis, M. Pfeffer, P. Jhund, A. S. Desai, J. J. V. McMurray, P. T. Ellinor, J. E. Ho, S. D. Solomon and J. W. Cunningham, Natural language processing to adjudicate heart failure hospitalizations in global clinical trials, *Circ Heart Fail* **18**, p. e012514 (2025).
14. C. Zheng, M. S. Lee, N. Bansal, A. S. Go, C. Chen, T. N. Harrison, D. Fan, A. Allen, E. Garcia, B. Lidgard, D. Singer and J. An, Identification of recurrent atrial fibrillation using natural language processing applied to electronic health records, *Eur Heart J Qual Care Clin Outcomes* **10**, 77 (2024).
15. X. Du, Z. Zhou, Y. Wang *et al.*, Performance and improvement strategies for adapting generative large language models for electronic health record applications: A systematic review, *Int J Med Inform* **205** (2025).
16. J. Vo, D. Weissenbacher, K. Reinier *et al.*, Leveraging large language models to identify in-hospital cardiac arrest, *JACC: Clinical Electrophysiology* (2026).
17. M. L. McHugh, Interrater reliability: the kappa statistic, *Biochem Med (Zagreb)* **22**, 276 (2012).

18. K. Reinier, H. S. Chugh, A. Uy-Evanado *et al.*, Observational study of sudden cardiac arrest risk (oscar): Rationale and design of an electronic health records cohort, *IJC Heart & Vasculature* **56**, p. 101614 (2025).
19. A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten and A. Vaughan, The llama 3 herd of models, *arXiv preprint arXiv:2407.21783* (2024).
20. J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le and D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, *Advances in neural information processing systems* **35**, 24824 (2022).
21. A. Yang, A. Li, B. Yang *et al.*, Qwen3 technical report (2025).
22. L. von Werra, Y. Belkada, L. Tunstall *et al.*, Trl: Transformer reinforcement learning `https://github.com/huggingface/trl`, (2020).
23. E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann and M. McDermott, Publicly available clinical bert embeddings, in *Proceedings of the 2nd clinical natural language processing workshop*, 2019.
24. M. Sokolova and G. Lapalme, A systematic analysis of performance measures for classification tasks, *Information Processing & Management* **45**, 427 (2009).
25. N. H. Shah, D. Entwistle and M. A. Pfeffer, Creation and adoption of large language models in medicine, *JAMA* **330**, 866 (09 2023).
26. S. Bedi, H. Cui, M. Fuentes *et al.*, Medhelm: Holistic evaluation of large language models for medical tasks (2025).
27. Y. S. Low, M. L. Jackson, R. J. Hyde *et al.*, Answering real-world clinical questions using large language model, retrieval-augmented generation, and agentic systems, *Digit Health.* **11** (2025).
28. D. Weissenbacher, M. Shabbir, I. M. Campbell, C. T. Berdahl and G. Gonzalez-Hernandez, Enhancing medical knowledge in large language models via supervised continued pretraining on clinical notes, *medRxiv* (2026).