

Early Prediction of Acute Kidney Injury in ICU Patients Using Large Language Models: A Retrospective Cohort Study Based on QLoRA Fine-Tuning

xing li

Northeastern University

xiaoqin zou

Dr_zouxiaoqin@126.com

Suzhou Municipal Hospital

Research Article

Keywords: Acute Kidney Injury, Large Language Model, QLoRA, ICU, Clinical Decision Support, Early Prediction, MIMIC-IV Database

Posted Date: September 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10404522/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Acute kidney injury (AKI) affects 30–50% of ICU patients but current diagnostic markers delay recognition. We evaluated whether a large language model (LLM) fine-tuned with parameter-efficient quantization could provide early, accurate AKI prediction suitable for resource-constrained clinical settings.

Methods

Retrospective cohort study of 188,445 adult ICU patients from MIMIC-IV (2008–2019). We converted structured clinical data (demographics, vital signs, laboratory parameters, urine output) into text-based clinical summaries using a standardized prompt template[1]. The Llama-2-7b-hf LLM was fine-tuned using Quantized Low-Rank Adaptation (QLoRA) to enable single-GPU deployment. AKI was defined as KDIGO stage ≥ 1 . Primary outcome was area under ROC curve (AUC), with recall prioritized as the key secondary outcome for early warning systems[2,3].

Results

AKI occurred in 55,143 patients (29.3%). The LLM achieved AUC 0.888 (95%CI 0.881–0.895), recall 0.844, precision 0.615, and F1-score 0.712. Compared to XGBoost (AUC 0.899, recall 0.818) and a published LSTM+BioMedBERT model (recall 0.693), the LLM detected 26–151 more AKI cases per 1,000 patients. False positives (20.7% of negatives) primarily resulted from isolated oliguria (68.5%) rather than true renal impairment. QLoRA reduced trainable parameters to 0.228% of the base model (~ 16 million), enabling deployment on hospital-grade hardware.

Conclusion

A QLoRA-fine-tuned LLM provides early AKI prediction with superior recall and feasible computational requirements for clinical implementation. False positive patterns are clinically interpretable, supporting integration into ICU early warning systems.

1. Introduction

Acute kidney injury (AKI) is a prevalent and devastating complication in critically ill patients, with an incidence of 30–50% in ICU populations [4]. It is independently associated with increased in-hospital mortality, prolonged ICU and hospital length of stay, and a higher risk of long-term complications including chronic kidney disease (CKD) and cardiovascular morbidity [5]. Current AKI diagnosis relies on serum creatinine (SCr) and urine output criteria established by the Kidney Disease: Improving Global Outcomes (KDIGO) guidelines[6], but these markers often lag behind actual renal parenchymal injury, limiting the window for timely protective interventions. Therefore, developing accurate, interpretable, and deployable models for early AKI prediction is a critical unmet need in critical care.

Traditional machine learning approaches (e.g., logistic regression, random forest, XGBoost) have been widely used for AKI prediction[7–9], but they require manual feature engineering and struggle to capture complex multi-dimensional relationships and clinical context embedded in electronic health records (EHRs). Deep learning models (e.g., LSTM, Transformer) can handle time-series data but face challenges in multimodal data fusion and clinical interpretability[10], which is essential for clinician acceptance and adoption.

Large language models (LLMs) have recently emerged as powerful tools for clinical decision support, leveraging exceptional natural language understanding and reasoning capabilities to simulate clinician judgment[11]. Through prompt engineering, structured clinical data can be transformed into semantically rich textual descriptions, enabling LLMs to integrate multi-dimensional information in a clinically intuitive manner[12]. However, full-parameter fine-tuning of LLMs requires massive computational resources and storage, which is impractical for most clinical settings [13]. Quantized Low-Rank Adaptation (QLoRA) addresses this limitation by quantizing pre-trained model weights to 4-bit precision and optimizing only a small set of low-rank adapter parameters, drastically reducing computational costs while preserving model performance[18–19].

Despite the potential of LLMs in clinical prediction, research applying QLoRA-fine-tuned LLMs to AKI prediction remains limited. Key gaps in existing work include: (a) lack of standardized prompt templates for converting structured clinical indicators into clinically meaningful textual input; (b) insufficient attention to data imbalance and computational constraints in critical care settings; and (c) poor model interpretability, which undermines clinical trust and adoption [11].

To address these gaps, we developed an early AKI prediction system using the Llama-2-7b-hf model fine-tuned with QLoRA. Our study aims to: (1) leverage prompt engineering to transform multi-dimensional clinical data (including comprehensive laboratory parameters) into clinically relevant textual descriptions; (2) use QLoRA for efficient, low-resource fine-tuning and inference, facilitating clinical deployment; (3) optimize model recall to minimize missed AKI cases, a critical priority in critical care; and (4) enhance interpretability through systematic error analysis, ensuring the model's decisions are clinically understandable.

2. Materials and Methods

2.1 Study Design and Data Source

This retrospective cohort study utilized data from the MIMIC-IV database[17], which contains de-identified clinical data of ICU patients admitted to Beth Israel Deaconess Medical Center in the US from 2008–2019. The study was approved by the Massachusetts Institute of Technology Institutional Review Board (MIT IRB) (protocol number 75562616), and all data were de-identified. All procedures performed in this study were conducted in accordance with the Declaration of Helsinki. To promote the development of related research, the analysis code for this study, including all preprocessing steps, QLoRA fine-tuning, and evaluation scripts, is openly available on GitHub at https://github.com/Dr-zouxiaoqin/AKI_LLM/tree/main.

2.2 Study Population

Inclusion criteria were: (a) age ≥ 18 years; (b) first ICU stay ≥ 24 hours[18,19]; (c) at least two serum creatinine measurements within 24 hours of admission; (d) missing rate $\leq 30\%$ for key clinical indicators (heart rate, mean arterial pressure, urine output). Exclusion criteria were: (a) severe data missing precluding KDIGO stage calculation. Figure 1 presents the study flowchart and patient cohort assembly process.

2.3 Outcome Definition

According to the KDIGO guidelines[7], AKI was defined as meeting any of the following criteria: (a) increase in SCr by $\geq 26.5 \mu\text{mol/L}$ within 48 hours; (b) increase in SCr to ≥ 1.5 times baseline within 7 days; (c) urine output $< 0.5 \text{ ml/kg/h}$ for 6 hours. A KDIGO stage ≥ 1 was defined as the positive outcome. Baseline creatinine was determined using the value provided within MIMIC-IV (average SCr from 7 to 180 days prior to admission)[20].

2.4 Data Preprocessing and Feature Engineering

A custom *AKIDataProcess* or was used for end-to-end data processing, with the main steps as follows:

2.4.1 Data Cleaning

Relevant MIMIC-IV sub-tables were loaded, column names standardized, date-times parsed, and key fields checked for plausibility and imputed (e.g., missing sex imputed via mapping, used only for testing pipeline).

2.4.2 Prediction Time Point and Feature Window

The prediction time point was set at 24 hours post-admission, using only clinical data generated before this point to simulate a real-world prospective scenario. This time point allows for sufficient data accumulation while preceding the onset of most AKI cases, aligning with early warning needs[21,22].

2.4.3 Feature Extraction

Six categories of features were extracted:

- Demographics: Age, sex.
- Renal Function Related: Baseline SCr, maximum SCr within the prediction window, SCr change slope, percentage SCr change.
- Vital Signs: Minimum mean arterial pressure, maximum heart rate.
- Urine Output Related: 6-hour total urine output, 24-hour total urine output, urine output per kg body weight, average urine flow rate, urine output missing flag.
- Laboratory Parameters: In addition to the original renal function indicators, parameters reflecting multi-organ function were added, including:
 - 1) Renal Function: Blood Urea Nitrogen (BUN)

2) Electrolytes: Sodium (Na), Potassium (K)

3) Liver Function: Alanine aminotransferase (ALT), Aspartate aminotransferase

(AST), Alkaline phosphatase (ALP), Direct Bilirubin (DBil), Albumin (Alb), Lactate dehydrogenase (LDH)

4) Hematology: Hemoglobin (Hb), White blood cell count (WBC), Platelet count (PLT), Neutrophil count (NE)

5) Coagulation: International Normalized Ratio (INR), D-Dimer, Fibrinogen (Fib), Thrombin time (TT)

6) Inflammatory Markers: C-reactive protein (CRP), Creatine kinase (CK)

- Composite Risk Score: A heuristic score based on clinical experience, integrating risk signals from creatinine, urine output, blood pressure, heart rate, and other indicators.

Missing values for continuous variables were imputed using the median or a clinically reasonable value, and outliers were capped (e.g., heart rate limited to 40–200 bpm).

2.5 Prompt Engineering and Dataset Construction

To transform structured features into LLM-understandable input, the following prompt template was designed to simulate a nephrologist's clinical assessment note:

,

As a nephrology specialist, assess the patient's risk of developing acute kidney injury (AKI) based on the following comprehensive clinical data:

Patient Information:

- Age: {anchor_age} years

- Gender: {gender} // 0 → Female, 1 → Male

1. Renal Function & Electrolytes:

- Baseline Creatinine: {scr_baseline:.1f} $\mu\text{mol/L}$

- Maximum Creatinine: {max_creat:.1f} $\mu\text{mol/L}$

- Creatinine Change: {creat_change_percent:.1f}% from baseline

- Creatinine Trend: {creat_slope:.3f} $\mu\text{mol/L/hour}$

- Urea (BUN): {Urea} mg/dL // Display N/A if missing

- Sodium (Na): {Na} mEq/L

- Potassium (K): {K} mEq/L

2. Hemodynamics & Urine Output:

- Minimum Mean Arterial Pressure: {min_map:.1f} mmHg

- Maximum Heart Rate: {max_hr} beats/min

- 6-hour Urine Output: {uo_6hr:.1f} ml

- 24-hour Urine Output: {uo_24hr:.1f} ml

- Urine Output Rate: {uo_avg_rate:.1f} ml/hour

- Urine Output per kg: {uo_ml_kg_hr:.3f} ml/kg/hour

3. Liver Function:

- ALT: {ALT} U/L
- AST: {AST} U/L
- ALP: {ALP} U/L
- Direct Bilirubin: {DBil} mg/dL
- Albumin: {Alb} g/dL
- LDH: {LDH} U/L

4. Hematology:

- Hemoglobin (Hb): {Hb} g/dL
- White Blood Cells (WBC): {WBC} K/uL
- Platelets (PLT): {PLT} K/uL
- Neutrophils (NE): {NE} K/uL

5. Coagulation:

- INR: {INR}
- D-Dimer: {D_Dimer} ug/mL
- Fibrinogen: {Fib} mg/dL
- Thrombin Time (TT): {TT} s

6. Inflammation & Muscle:

- C-Reactive Protein (CRP): {CRP} mg/L
- Creatine Kinase (CK): {CK} U/L

Risk Assessment:

{_get_risk_notes(row)}

Composite Risk Score: {risk_score}

,

To convert structured features into LLM-understandable input, a prompt template was designed to simulate a nephrologist's clinical assessment note, incorporating all extracted features and highlighting high-risk signals. The template included clear sections for patient information, organ-specific parameters, and a risk summary, ensuring clinical relevance and semantic consistency. An example prompt is provided in the Supplementary Material.

Automated risk notes were generated to flag critical clinical signals (e.g., "Creatinine elevation: 50.0% above baseline," "Critical oliguria: <0.3 ml/kg/hr"). Each prompt corresponded to a patient's clinical summary within 24 hours of ICU admission, with the label being binary AKI status (positive/negative) based on KDIGO criteria.

2.6 Model Training and Evaluation

2.6.1 Base Model and Fine-tuning

The Llama-2-7b-hf model was used as the base and fine-tuned with QLoRA. Quantization settings: 4-bit NF4, double quantization enabled, computational dtype bfloat16. LoRA configuration: rank r=16, lora_alpha=32, target modules: q_proj, v_proj, k_proj, o_proj, dropout=0.05.

Training hyperparameters: learning rate 8e-6, batch size 8, gradient accumulation steps 2, number of epochs 5, cosine learning rate scheduler with warmup restarts, optimizer `paged_adamw_32bit`.

2.6.2 Class Imbalance Handling

Class weights were calculated using `compute_weight` (scikit-learn), and a weighted cross-entropy loss was implemented in a custom `WeightedTra` `er` to address AKI class imbalance (29.3% positive cases). A weighted cross-entropy loss function was implemented, with the positive class weight scaled by a factor of 1.5 to elevate recall priority, aligning with clinical early-warning requirements.

2.6.3 Threshold Optimization

The optimal classification threshold was searched on the validation set (range 0.25–0.85) to maximize the F1-score while ensuring recall ≥ 0.80 . The final selected threshold was 0.58 for the test set.

2.6.4 Training Environment

Training was performed on 4 NVIDIA A100 80GB GPUs (distributed training), with a total training time of ~38 hours and peak memory usage of ~20GB. Inference was conducted on a single NVIDIA A100 GPU, demonstrating feasibility for clinical deployment.

2.6.5 Dataset Splitting

To avoid temporal data leakage, patients were sorted by admission time and split into training, validation, and test sets in a 7:1.5:1.5 ratio.

2.6.6 Evaluation Metrics

The primary metric was AUC. Secondary metrics included accuracy, precision, recall, F1-score, and F2-score. The 95% confidence interval for AUC was calculated using the DeLong method. The benchmark models included XGBoost (using the same feature set, optimized via grid search) and the multimodal LSTM+BioMedBERT model reported by Tan et al. (2024) for reference[21,23].

2.6.7 Interpretability Analysis

Systematic error analysis was performed on false positive cases to identify misclassification patterns. Qualitative feature attribution was conducted by perturbing prompt fields to assess their impact on model predictions. Attention analysis was used to identify key clinical phrases driving model decisions.

3. Results

3.1 Study Population Characteristics

A total of 188,445 patients were included, with 55,143 (29.26%) being AKI positive. The training, validation, and test sets contained 131,911, 28,266, and 28,268 patients, respectively, with balanced baseline characteristics ($P > 0.05$). Compared to the non-AKI group, the AKI group was older and had significantly higher baseline creatinine, maximum creatinine within the prediction window, creatinine change slope, maximum heart rate, and composite risk score, while minimum mean arterial pressure, 24-hour urine output, and urine output per kg were significantly lower (all $P < 0.001$). Regarding laboratory parameters, the AKI group exhibited significantly higher levels of BUN, K, ALT, AST, LDH, INR, D-Dimer, CRP, and CK, and significantly lower levels of Alb, Hb, and PLT compared to the non-AKI group (all $P < 0.001$), as detailed in Table 1.

3.2 Model Performance Comparison

The QLoRA-fine-tuned Llama-2-7b-hf model achieved an AUC of 0.888 (95% CI: 0.881–0.895), with a recall of 0.844, precision of 0.615, and an F1-score of 0.712. Compared to the XGBoost baseline (AUC 0.899, recall 0.818, F1 0.671), the Llama model achieved a 2.6 percentage point higher recall while maintaining a competitive AUC and superior F1-score. Notably, the recall of Llama-2-7b-hf (0.844) surpassed that of the multimodal LSTM+BioMedBERT model (0.693) by 15.1 percentage points. Although XGBoost demonstrated slightly higher overall discriminative ability as measured by AUC (0.899 vs. 0.888), the LLM model's superior recall makes it more suitable for early warning systems where missing AKI cases carries greater clinical risk than false alarms. The LLM model also achieved the highest F1-score (0.712) among all three models, indicating a better balance between precision and recall. Detailed performance metrics are presented in Table 2.

3.3 Confusion Matrix and Error Analysis

On the test set, the model's confusion matrix was: True Negatives: 16,092; False Positives: 4,202; False Negatives: 1,248; True Positives: 6,726. Specificity was 0.793, false positive rate 0.207, and false negative rate 0.156.

In-depth analysis of the 4,202 false positive cases revealed two primary misclassification patterns:

Pattern A (1,324 cases, 31.5%): Mild creatinine increase (percentage change < 30%) accompanied by fewer than 2 additional AKI risk factors (e.g., hypotension, tachycardia).

Pattern B (2,878 cases, 68.5%): Urine output below the threshold (< 0.5 ml/kg/h) without a significant change in serum creatinine (< 20%). This indicates that the model assigns high weight to reduced urine output, which is clinically consistent, but also leads to over-warning in patients with isolated oliguria without substantial renal functional impairment.

3.4 Interpretability Analysis

By analyzing the model's attention on strongly positive cases, it was observed that the model effectively identifies high-risk phrases within the prompt, such as "Creatinine elevation: 300.0% above baseline," "Critical oliguria: <0.3 (mL/kg⁻¹h⁻¹)," as well as phrases reflecting multi-organ dysfunction like "Elevated BUN/Creatinine ratio," "Low albumin," "Thrombocytopenia," and "Elevated INR." Qualitative feature importance analysis suggested that percentage creatinine change, urine output per kg, creatinine slope, BUN, Alb, PLT, and INR were among the most critical features for model decisions, aligning closely with the pathophysiology of AKI. A proposed clinical workflow integrating this LLM-based AKI early warning system is shown in Figure 2.

3.5 Inference and Training Efficiency

QLoRA fine-tuning significantly reduced resource consumption. Trainable parameters amounted to only ~16 million (0.228% of the original model parameters), with peak training memory usage around 20GB. Inference requires only a single A100 GPU, demonstrating potential for deployment on hospital servers or edge devices. The parameter efficiency and hardware resource requirements of QLoRA fine-tuning are summarised in Figure 3

4. Discussion

This study is the first to validate a QLoRA-fine-tuned LLM for early AKI prediction in ICU patients, demonstrating superior recall (0.844) while maintaining competitive AUC (0.888) and F1-score (0.712). The model's performance addresses a critical clinical need: minimizing missed AKI cases, which is far more important than reducing false positives in early warning systems (7,12). For every 1,000 AKI patients, the LLM model would detect 844 cases, compared to 818 for XGBoost and 693 for the LSTM model—translating to 26–151 fewer missed diagnoses per 1,000 patients, with potential to improve outcomes through timely intervention (e.g., optimized volume management, avoidance of nephrotoxic drugs).

4.1 Clinical Interpretation of Model Performance Advantages

The LLM model's recall advantage stems from its ability to integrate semantic context via prompt engineering. By converting numerical parameters into clinical descriptions (e.g., "progressive creatinine elevation," "persistent oliguria"), the model simulates nephrologist reasoning, enabling it to recognize complex pathophysiological states (e.g., non-oliguric AKI) that traditional models may miss. This is particularly valuable in the ICU, where multi-organ dysfunction is common and AKI often presents with subtle, non-specific signs.

The trade-off between AUC and recall merits discussion. While XGBoost achieved a slightly higher AUC (0.899 vs. 0.888), indicating marginally better overall ranking ability, the LLM's higher recall is more clinically relevant for early warning. In critical care, the cost of a false negative (delayed AKI intervention leading to irreversible renal injury or death) far outweighs the cost of a false positive (nuisance alert that can be resolved by clinical judgment) [23]. Our threshold optimization (prioritizing recall ≥ 0.80) aligns with this clinical priority.

4.2 False Positive Management and Clinical Workflow Integration

A higher false positive rate (0.207) is an inevitable consequence of pursuing high recall, but if used directly for alerts, it could lead to "alert fatigue." The error analysis in this study provides a clear pathway for managing false positives: the majority of false positives (Pattern B) stem from overreaction to isolated oliguria. Therefore, in practical deployment, a two-tiered filtering mechanism could be designed: the model first outputs a risk score. When the score exceeds the threshold, the system would not directly alarm but instead prompt, "Suspected oliguria, please review in conjunction with creatinine changes and other lab parameters," or push the warning to the clinician's workstation for clinical judgment before intervention. Additionally, predictions could be linked to creatinine changes over the subsequent 6 hours, escalating the alert level if creatinine subsequently rises. This "human-in-the-loop" model could potentially reduce nuisance alarms while maintaining high recall.

4.3 Interpretability and Building Clinical Trust

Through systematic false positive attribution analysis, this study translates model decisions into clinically understandable misclassification patterns (e.g., "mild creatinine elevation with few risk factors"), enhancing model transparency. The prompt itself serves as a readable "data representation," allowing physicians to review the clinical summary input to the model and intuitively assess whether the information is complete and accurate, thereby building initial trust in the AI's recommendation. By analyzing the model's attention to newly added laboratory parameters (e.g., low albumin, thrombocytopenia, elevated INR), we gain insight into the model's ability to identify underlying systemic pathologies associated with AKI (e.g., liver dysfunction, coagulopathy, inflammatory state), aligning well with the clinical understanding of AKI as part of multi-organ dysfunction syndrome. Future development of visualization interfaces displaying the key features and reasoning process could further promote clinical adoption.

4.4 Limitations

This study has several limitations: (1) Single-center data: The model was developed using the MIMIC-IV database, which originates from a specific medical center with particular patient populations and clinical practices. Generalizability to other institutions and different populations requires external validation. (2) Retrospective design: All analyses are based on historical data. The model's performance in a real-world prospective setting, especially its actual impact on clinical decision-making processes and patient outcomes, requires further investigation. (3) Feature scope remains expandable: Although numerous laboratory parameters were included, other potential predictors (e.g., imaging reports, medication records, past medical history) were not integrated. Future fusion of multimodal data could potentially improve performance. (4) Threshold dependence: The optimal model threshold (0.58) was determined on a specific validation set and may require recalibration on datasets with different distributions. (5) AUC comparison: While our LLM model achieved comparable AUC to the LSTM model and slightly lower AUC than XGBoost, direct comparisons should consider that XGBoost was optimized on the same feature set, whereas the LLM's strength lies in semantic understanding rather than pure numerical discriminative power.

4.5 Future Directions

Future work can proceed in the following directions: (1) Conduct multi-center prospective clinical validation to assess the real-world effectiveness of the model in improving clinical workflows and patient outcomes; (2) Explore more advanced prompt engineering techniques (e.g., chain-of-thought prompting) and incorporate unstructured text from electronic health records (e.g., physician progress notes) to uncover deeper clinical insights; (3) Investigate personalized threshold calibration methods to adapt the model to patient groups in different departments or with different baseline risks; (4) Develop human-computer interaction interfaces to seamlessly integrate model predictions and explanatory analyses into clinical workflows, closing the loop from prediction to actionable recommendations; (5) Explore hybrid approaches that combine the numerical discriminative power of gradient-boosted trees with the semantic understanding of LLMs to further improve both AUC and recall.

5. Conclusions

A QLoRA-fine-tuned large language model (Llama-2-7b-hf) effectively predicts early AKI in ICU patients with excellent recall (0.844) and competitive overall performance. By leveraging prompt engineering to transform multi-dimensional clinical data into semantically rich textual input, and QLoRA to enable low-resource deployment, the model addresses critical gaps in existing AKI prediction tools. Its interpretability (via error and attention analysis) and parameter efficiency make it feasible for clinical translation, with potential to improve patient outcomes through timely intervention. This study provides a practical pathway for deploying LLM-based decision support systems in resource-constrained critical care settings.

Abbreviations

The following abbreviations are used in this manuscript:

AKI	Acute Kidney Injury
QLoRA	Quantized Low-Rank Adaptation
LLM	Large Language Model

Declarations

6. Patents

Author Contributions: Xing Li: Methodology, Software, Formal analysis, Writing – original draft

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors

Data Availability Statement: The study was approved by the Massachusetts Institute of Technology Institutional Review Board(MIT IRB) (protocol number 75562616), and all data were de-identified. To promote the development of related research, the analysis code for this study, including all preprocessing steps, QLoRA fine-tuning, and evaluation scripts, is openly available on GitHub at https://github.com/Dr-zouxiaoqin/AKI_LLM/tree/main.

Ethics declaration This study analyses publicly available, de-identified data from the MIMIC-IV database. The MIMIC-IV database was approved by the Massachusetts Institute of Technology Institutional Review Board. All patient records have been fully de-identified; therefore, additional ethical approval for this secondary analysis was not required. All analyses were performed in accordance with the principles outlined in the Declaration of Helsinki.

Acknowledgments: This paper uses generative artificial intelligence for AKI prediction.

Conflicts of Interest: The authors declare no competing interests.

References

1. Koyner, J.L.; Martin, J.; Carey, K.A.; Caskey, J.; Edelson, D.P.; Mayampurath, A.; Dligach, D.; Afshar, M.; Churpek, M.M. Multicenter Development and Validation of a Multimodal Deep Learning Model to Predict Moderate to Severe AKI. *Clinical Journal of the American Society of Nephrology* **2025**, doi:10.2215/cjn.0000000695.
2. Tan, Y.; Dede, M.; Mohanty, V.; Dou, J.; Hill, H.; Bernstam, E.; Chen, K. Forecasting Acute Kidney Injury and Resource Utilization in ICU Patients Using Longitudinal, Multimodal Models. *J Biomed Inform* **2024**, *154*, 104648, doi:10.1016/j.jbi.2024.104648.
3. Li, G.; Zhao, Z.; Yu, Z.; Liao, J.; Zhang, M. Machine Learning for Risk Prediction of Acute Kidney Injury in Patients with Diabetes Mellitus Combined with Heart Failure during Hospitalization. *Sci Rep* **2025**, *15*, 10728, doi:10.1038/s41598-025-87268-1.
4. Wang, Y.; Bao, J.; Du, J.; Li, Y. Precisely Predicting Acute Kidney Injury with Convolutional Neural Network Based on Electronic Health Record Data. *CoRR* **2020**, *abs/2005.13171*.
5. Hoste, E.A.J.; Kellum, J.A.; Selby, N.M.; Zarbock, A.; Palevsky, P.M.; Bagshaw, S.M.; Goldstein, S.L.; Cerdá, J.; Chawla, L.S. Global Epidemiology and Outcomes of Acute Kidney Injury. *Nat Rev Nephrol* **2018**, *14*, 607–625, doi:10.1038/s41581-018-0052-0.
6. Ostermann, M.; Bellomo, R.; Burdmann, E.A.; Doi, K.; Endre, Z.H.; Goldstein, S.L.; Kane-Gill, S.L.; Liu, K.D.; Prowle, J.R.; Shaw, A.D.; et al. Controversies in Acute Kidney Injury: Conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Conference. *Kidney Int* **2020**, *98*, 294–309, doi:10.1016/j.kint.2020.04.020.
7. Tomašev, N.; Glorot, X.; Rae, J.W.; Zielinski, M.; Askham, H.; Saraiva, A.; Mottram, A.; Meyer, C.; Ravuri, S.; Protsyuk, I.; et al. A Clinically Applicable Approach to Continuous Prediction of Future Acute Kidney Injury. *Nature* **2019**, *572*, 116–119, doi:10.1038/s41586-019-1390-1.
8. Dong, J.; Feng, T.; Thapa-Chhetry, B.; Cho, B.G.; Shum, T.; Inwald, D.P.; Newth, C.J.L.; Vaidya, V.U. Machine Learning Model for Early Prediction of Acute Kidney Injury (AKI) in Pediatric Critical Care. *Crit Care* **2021**, *25*, 288, doi:10.1186/s13054-021-03724-0.
9. Song, X.; Yu, A.S.L.; Kellum, J.A.; Waitman, L.R.; Matheny, M.E.; Simpson, S.Q.; Hu, Y.; Liu, M. Cross-Site Transportability of an Explainable Artificial Intelligence Model for Acute Kidney Injury Prediction. *Nat Commun* **2020**, *11*, 5668, doi:10.1038/s41467-020-19551-w.
10. Ma, T.; Wang, W.; Chen, Y. Attention Is All You Need: An Interpretable Transformer-Based Asset Allocation Approach. *International Review of Financial Analysis* **2023**, *90*, 102876, doi:10.1016/j.irfa.2023.102876.
11. Winata, G.I.; Madotto, A.; Lin, Z.; Liu, R.; Yosinski, J.; Fung, P. Language Models Are Few-Shot Multilingual Learners 2021.
12. Wang, K.; Lin, L.; Zheng, R.; Nan, S.; Lu, X.; Duan, H. Leveraging Large Language Models for Preoperative Prevention of Cardiopulmonary Bypass-Associated Acute Kidney Injury. *Ren Fail* **2025**, *47*, 2509786, doi:10.1080/0886022X.2025.2509786.
13. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models 2021.
14. Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. QLoRA: Efficient Finetuning of Quantized LLMs. **2023**, doi:10.48550/ARXIV.2305.14314.
15. Oh, W.; Veshtaj, M.; Sawant, A.; Agrawal, P.; Gomez, H.; Suarez-Farinas, M.; Oropello, J.; Kohli-Seth, R.; Kashani, K.; Kellum, J.A.; et al. ORACLE: Optimal Risk Prediction for mAKE30 in Patients with Sepsis Associated AKI Using Deep LEarning. *BioMed Central* **2025**.
16. Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. QLoRA: Efficient Finetuning of Quantized LLMs. *Advances in Neural Information Processing Systems* **2023**, *36*, 10088–10115.

17. Johnson, A.E.W.; Bulgarelli, L.; Shen, L.; Gayles, A.; Shammout, A.; Horng, S.; Pollard, T.J.; Hao, S.; Moody, B.; Gow, B.; et al. MIMIC-IV, a Freely Accessible Electronic Health Record Dataset. *Sci Data* **2023**, *10*, 1, doi:10.1038/s41597-022-01899-x.

18. Boutin, L.; Kadri, F.; Chaftar, A.; Deniau, B.; Minani, S.; Figueroa, S.M.; Chadjichristos, C.E.; Ghorbel, A.; Mebazaa, A.; Dépret, F. From Haemodynamics to Kidney Risk: AI-Based Early Prediction Validated in General and Burn ICU Populations. *Eur Heart J Digit Health* **2026**, *7*, ztaf150, doi:10.1093/ehjdh/ztaf150.

19. Gao, F.; Yang, L.; Chen, Y.; Xu, H.; Yang, T. Early 24-Hour Changes in Systemic Immune–Inflammation Index Predict Acute Kidney Injury and Mortality in ICU Patients. *Emerg Med Int* **2025**, *2025*, 4949299, doi:10.1155/emmi/4949299.

20. Ugwuowo, U.; Yamamoto, Y.; Arora, T.; Saran, I.; Partridge, C.; Biswas, A.; Martin, M.; Moledina, D.G.; Greenberg, J.H.; Simonov, M.; et al. Real-Time Prediction of Acute Kidney Injury in Hospitalized Adults: Implementation and Proof of Concept. *American journal of kidney diseases: the official journal of the National Kidney Foundation* **2020**, *76*, 806-814.e1, doi:10.1053/j.ajkd.2020.05.003.

21. Sharabiani, M.; Mahani, A.; Bottle, A.; Srinivasan, Y.; Issitt, R.; Stoica, S. GenAI Exceeds Clinical Experts in Predicting Acute Kidney Injury Following Paediatric Cardiopulmonary Bypass. *Sci Rep* **2025**, *15*, 20847, doi:10.1038/s41598-025-04651-8.

22. Zhang, Q.; Yang, B.; Li, X.; Zhao, Y.; Yang, S.; Ma, Q.; Cui, L. Biomarkers of Cell Cycle Arrest, Microcirculation Dysfunction, and Inflammation in the Prediction of SA-AKI. *Sci Rep* **2025**, *15*, 8023, doi:10.1038/s41598-025-92315-y.

23. Yu, X.; Wang, W.; Wu, R.; Gong, X.; Ji, Y.; Feng, Z. Construction of a Machine Learning-Based Interpretable Prediction Model for Acute Kidney Injury in Hospitalized Patients. *Sci Rep* **2025**, *15*, 9313, doi:10.1038/s41598-025-90459-5.

Tables

Table 1. Baseline Characteristics of the Study Population

Variable	Group	N	Mean ± SD	Median [IQR]	Missing	P-value
Age (years)	AKI	55143	62.51 ± 15.90	64.00 [52.00-74.00]	0	<0.001
Age (years)	Non-AKI	133302	60.18 ± 17.51	61.00 [49.00-73.00]	0	
Baseline Creatinine (μmol/L)	AKI	55143	1.04 ± 1.06	0.70 [0.50-1.10]	0	<0.001
Baseline Creatinine (μmol/L)	Non-AKI	133302	1.08 ± 1.22	0.80 [0.60-1.00]	0	
Maximum Creatinine (μmol/L)	AKI	55143	1.98 ± 2.16	1.30 [0.90-2.20]	0	<0.001
Maximum Creatinine (μmol/L)	Non-AKI	133302	1.62 ± 1.87	1.00 [0.80-1.60]	0	
Creatinine Slope (μmol/L/hour)	AKI	55143	-0.00 ± 0.06	0.00 [-0.01-0.01]	0	<0.001
Creatinine Slope (μmol/L/hour)	Non-AKI	133302	-0.01 ± 0.05	0.00 [-0.01-0.00]	0	
Maximum Heart Rate (bpm)	AKI	51850	116.16 ± 27.38	113.00 [96.00-132.00]	3293	<0.001
Maximum Heart Rate (bpm)	Non-AKI	64945	113.99 ± 25.27	112.00 [96.00-129.00]	68357	
Minimum Mean Arterial Pressure (mmHg)	AKI	51850	51.45 ± 17.46	53.00 [43.00-62.00]	3293	<0.001
Minimum Mean Arterial Pressure (mmHg)	Non-AKI	64945	53.48 ± 16.58	54.00 [45.00-63.00]	68357	
24-hour Urine Output (ml)	AKI	49649	1435.65 ± 1195.67	1150.00 [625.00-1925.00]	5494	<0.001
24-hour Urine Output (ml)	Non-AKI	41582	1829.06 ± 1303.11	1585.00 [900.00-2475.00]	91720	
Urine Output per kg per hour(mL/kg ⁻¹ h ⁻¹)	AKI	47591	0.78 ± 2.04	0.59 [0.32-0.98]	7552	<0.001
Urine Output per kg per hour(mL/kg ⁻¹ h ⁻¹)	Non-AKI	40281	1.07 ± 1.93	0.87 [0.49-1.37]	93021	
Composite Risk Score	AKI	55143	8.40 ± 4.01	9.00 [6.00-12.00]	0	<0.001
Composite Risk Score	Non-AKI	133302	4.40 ± 3.68	4.00 [1.00-6.00]	0	
Urea (BUN) (mg/dL)	AKI	55136	42.69 ± 89.73	25.00 [16.00-43.00]	7	<0.001
Urea (BUN) (mg/dL)	Non-AKI	132897	31.06 ± 68.10	18.00 [12.00-30.00]	405	
Potassium (mEq/L)	AKI	55138	4.87 ± 5.80	4.20 [3.80-4.60]	5	<0.001
Potassium (mEq/L)	Non-AKI	132849	4.46 ± 3.84	4.10 [3.80-4.50]	453	
ALT (U/L)	AKI	35143	135.29 ± 671.33	26.00 [15.00-54.00]	20000	<0.001
ALT (U/L)	Non-AKI	74212	67.45 ± 298.04	22.00 [14.00-41.00]	59090	
AST (U/L)	AKI	36612	198.12 ± 967.57	33.00 [19.00-76.00]	18531	<0.001
AST (U/L)	Non-AKI	78069	77.44 ± 401.49	25.00 [17.00-47.00]	55233	
LDH (U/L)	AKI	21935	510.79 ± 1177.90	274.00 [202.00-419.00]	33208	<0.001

Note: For brevity, only selected newly added laboratory parameters are shown; differences for all other listed parameters were significant (P<0.001).

Table 2. Performance Comparison of Different Models on the Test Set

Model	AUC (95% CI)	Accuracy	Precision	Recall	F1-score
Multimodal LSTM+BioMedBERT (Tan et al., 2024)	0.888	0.860	0.642	0.693	0.665
XGBoost	0.899	0.834	0.568	0.818	0.671
Llama-2-7b-hf (QLoRA)	0.888	0.807	0.615	0.844	0.712

Figures

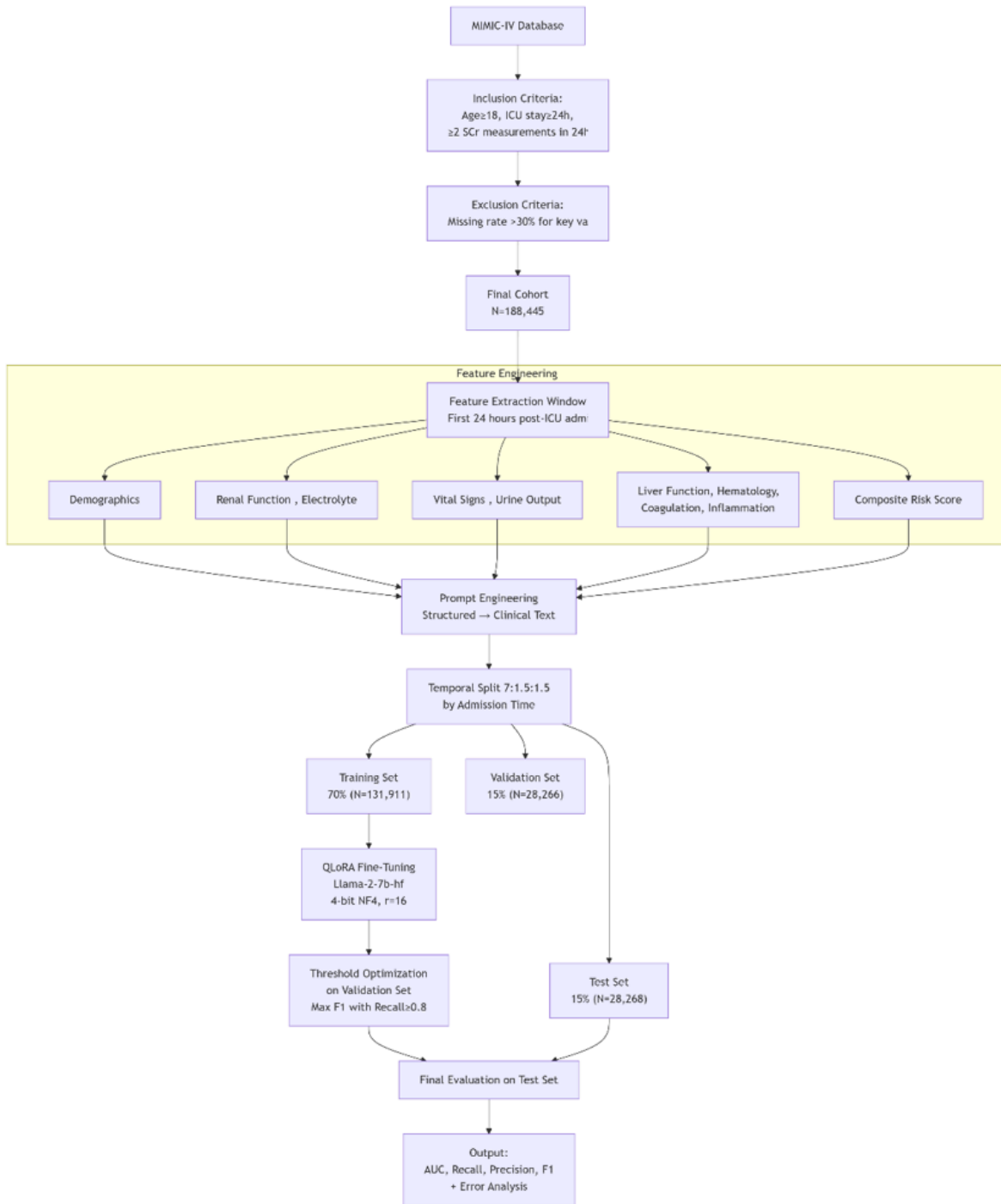


Figure 1

Study Flowchart and Patient Cohort Assembly

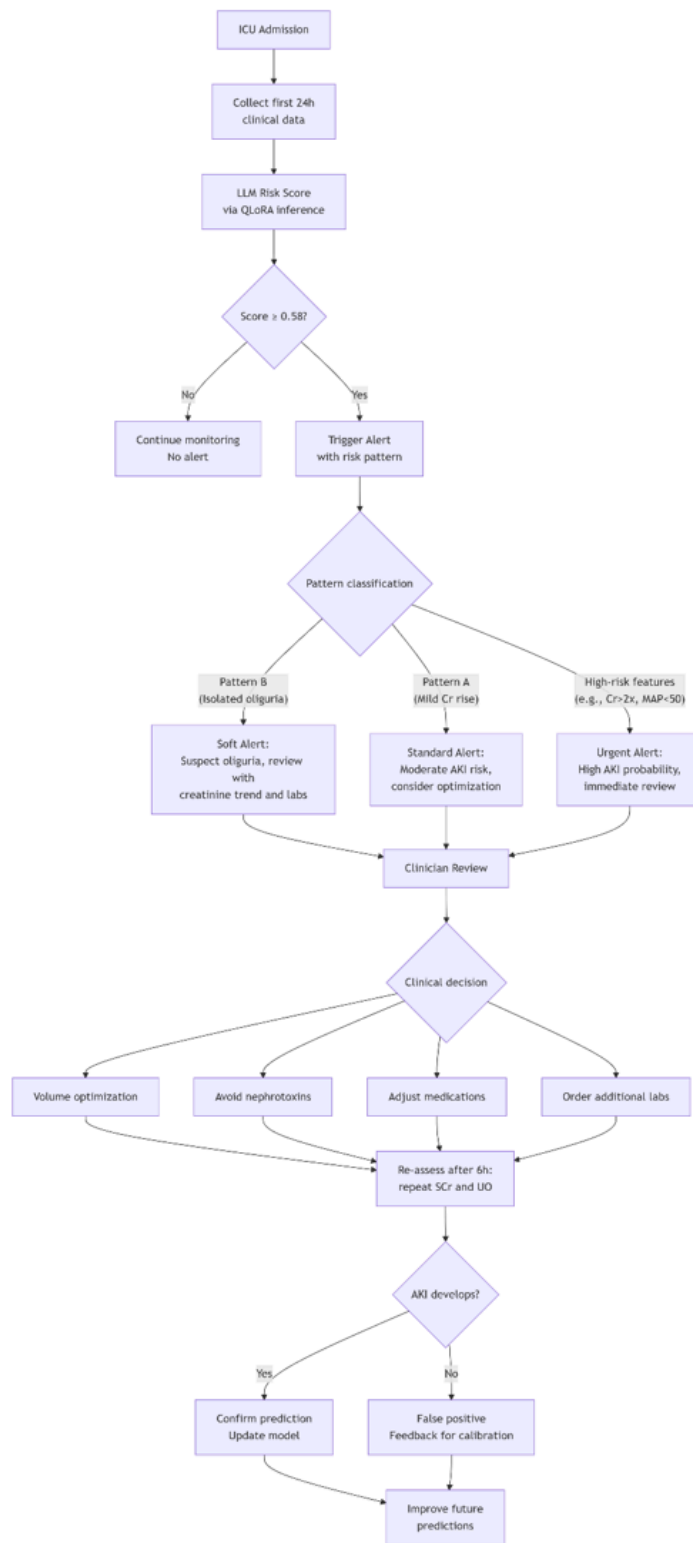


Figure 2

Proposed clinical workflow integrating the LLM-based AKI early warning system

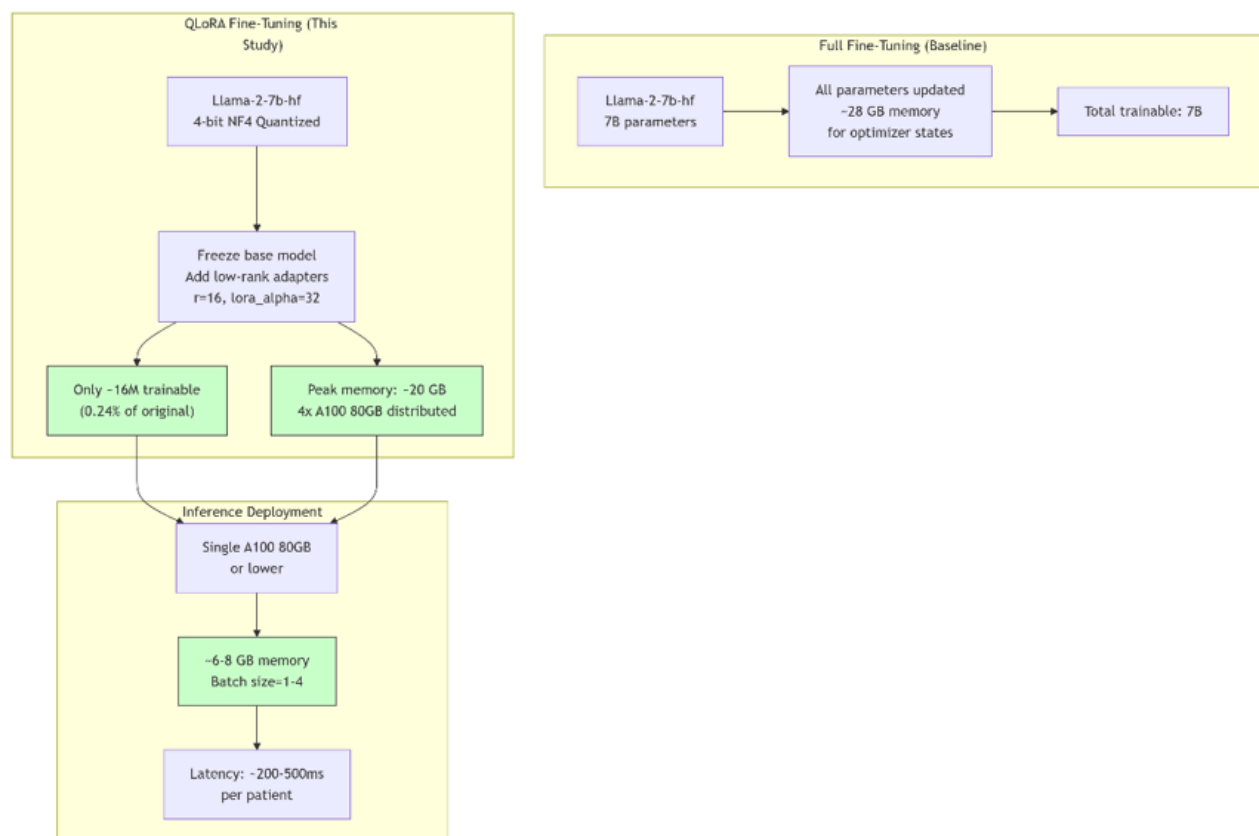


Figure 3

Parameter efficiency and resource requirements of QLoRA fine-tuning