nature portfolio

Peer Review File

Large Language Model Driven Multicenter Prediction and Explainable Risk Attribution of Acute Kidney Injury

Corresponding Author: Professor Xizi Zheng

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Xu et al. introduces a dual-model LLM framework (AKI-PM and AKI-RAM) designed for real-world AKI prediction and clinically interpretable risk attribution, built through extensive pretraining, fine-tuning, and multimodal alignment on EHR-derived text.
The conceptual framing is compelling, but deployment in actual clinical environments is not currently feasible. Achieving REAL-TIME AKI prediction with actionable interpretability would require deep integration of these models into EHR systems, a capability that remains under active development given unresolved issues around data governance, transparency, privacy, regulatory constraints, and the exceptionally high-performance standards required in healthcare. Their performance on curated and well-organized EHR datasets does not necessarily reflect how they would perform when fully integrated into a LIVE EHR system, where additional operational, technical, and clinical factors (complicated clinical information, such as different types of notes, labs, and images) can substantially influence model behavior. As a result, the present work has limited practical applicability at this stage.
Furthermore, several methodological gaps need to be addressed before firm conclusions can be drawn:
1. Datasets such as EBM-text, PUB-test, and Public-Note would be expected to be in English. What language was used across the internal and external datasets, as well as in the prompts—English or Chinese?
2. The comparison against generalist LLMs remains methodologically imbalanced since these models were not fine-tuned for structured EHR prediction, whereas AKI-PM was extensively trained on domain-specific clinical data. The reported extreme failure of generalist LLMs (specificity 0.00–0.25 in the main paper) likely reflects mismatched task design, not true inferiority.
3. Regarding AKI-RAM, does the model restrict its attention to the examples included in the prompt? For instance, when asked to explain intervenable risk factors, does it limit its response to risk elements that clinicians could realistically modify?
4. Regarding the df2text pipeline (Supplementary Method Section 3.1 and Supplementary Figure 5), several critical details remain unaddressed:
a. df2text does not address time-interval weighting or encoding of irregular sampling beyond linear text ordering. How does the model account for irregular sampling intervals beyond linear ordering?
b. The truncation to 16 rows (Supplementary Method Section 3.3) may introduce information loss. What is the effect of truncating to 16 rows on prediction accuracy and how many patients experience significant loss of historical data?
c. The supplement demonstrates a 0.94 to 0.81 AUC drop (Accuracy 0.90 vs 0.62, Specificity 0.92 vs 0.58) when converting df2text to df2string (Supplementary Table 8), yet does not explain why string formatting is so detrimental.
5. The pretraining corpus (AKICorpus) includes institutional EHR-Note data drawn from the same center used for fine-tuning and internal validation, raising the possibility of subtle leakage of institutional patterns, coding habits, or lexical conventions.
6. The deduplication thresholds (α=0.7, β=0.88) may allow retention of semantically similar clinical trajectories. Whether were near-duplicate cases in the EHR corpus inspected or filtered?
7. The largest component is drawn from MIMIC-IV and eICU—nearly entirely ICU-heavy datasets, yet the model systematically performs worse in ICU populations across all centers (Supplementary Tables 4–7). How to explain this asymmetry and potential disparities?
8. The manuscript highlights few-shot improvement for external centers, but the supplement does not describe: How those 1,000 samples were selected or distributed across AKI vs non-AKI? Whether prompting, calibration, or underlying weights were modified? Whether smaller adaptation sets (e.g., 10–200 samples) were tested?
9. The rejected responses were generated by DeepSeek-R1. Whether rejected responses might systematically bias the model toward particular reasoning patterns rather than toward universally valid clinical logic?

10. Supplementary Method Section 2 describes deterministic prompt templates, but the inference section (Section 6) notes that invalid outputs require repeated attempts. Please report the failure rate across the full validation set and consider implications for deployment stability.
11. Supplementary Method Section 6 indicates that inference may require up to three passes before a valid response is generated. Whether this affects real-time prediction feasibility and how output validation is handled in live settings?
12. Supplementary Figure 3 provides a sample AKI-RAM output, but interpretation concerns remain: The attribution schema is heavily prompt-driven rather than model-internal. It is not possible to determine whether AKI-RAM identifies genuine causal mechanisms or merely reorders evidence extracted from df2text. Evaluation of AKI-RAM also lacks inter-rater agreement and external raters to mitigate institutional bias.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
This is a nice attempt to develop a model for AKI prediction across all hospitalized patients. The model works fine, and most importantly, it provides modifiable factors for potential intervention. Though not prospectively validated in the real world, it is still worth being published to pave the way for future validation. I am not an expert in LLM, but I have several comments to further improve the manuscript based on my expertise in clinical studies.
1. The previous studies were both conducted in enriched populations, so this study is a major step forward. This should be mentioned in the Introduction.
2. Line 331-332，how the authors define "nephrologist";
3. Ethical approval should be obtained in all the participating sites, including those for external validation.
4. How the authors define modifiable factors should be detailed;
5. What is the weight of the two modifiable factors, "insufficient fluid" and "new infection" in the prediction? Can they be quantified? If the weight is too low, the clinical benefits may be just marginal with extra huamn resource cost.
6. What does the factor "new-onset infection" mean? or what is it based on? If it is based on medical records, I don't think it is really "modifiable". Once the diagnosis is made by the physician, treatment usually ensues. If it is based on microbiology findings and lab results, please also report its consistency with clinical diagnosis.
7. How will the model perform in an enriched population with high risk? This can be interesting. Like those defined by non-modifiable factors such as advanced age;
This is an interesting study that developed a mode for a very specific purpose. However, predicting AKI is not like predicting MI, there are very few things that can be done even if you know the patients are at risk. The real clinical gain from the model needs to be tested in a rigorously designed trial.
8.

(Remarks on code availability)
I am not an expert on that.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have addressed all my concerns.

(Remarks on code availability)
The results of the paper are reproducible, and the code is a usable resource for the community.

Reviewer #3

(Remarks to the Author)
Thanks to the authors for their constructive responses and revisions. I have no further ideas.

(Remarks on code availability)

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

Xu et al. introduces a dual-model LLM framework (AKI-PM and AKI-RAM) designed for real-world AKI prediction and clinically interpretable risk attribution, built through extensive pretraining, fine-tuning, and multimodal alignment on EHR-derived text.
The conceptual framing is compelling, but deployment in actual clinical environments is not currently feasible. Achieving REAL-TIME AKI prediction with actionable interpretability would require deep integration of these models into EHR systems, a capability that remains under active development given unresolved issues around data governance, transparency, privacy, regulatory constraints, and the exceptionally high-performance standards required in healthcare. Their performance on curated and well-organized EHR datasets does not necessarily reflect how they would perform when fully integrated into a LIVE EHR system, where additional operational, technical, and clinical factors (complicated clinical information, such as different types of notes, labs, and images) can substantially influence model behavior. As a result, the present work has limited practical applicability at this stage.
Response: We appreciate the reviewer's important practical concerns. We fully agree that live EHR integration involves substantial challenges related to data governance, privacy, regulatory compliance, and interoperability, and we have revised the manuscript to acknowledge them more explicitly.
Nevertheless, we would like to offer several clarifications regarding the feasibility claimed in our work. Our use of "real-time" refers to generating AKI risk predictions at 6-hour intervals, with a mean inference latency of only 0.23 seconds per case. This represents a technically meaningful form of real-time operation, reflecting the model's potential for timely clinical decision support, as demonstrated in analogous published LLM-based clinical prediction systems (e.g., Shashikumar et al., NPJ Digit Med, 2025). We have clarified this definition in the manuscript. In this retrospective, multi-center development and validation study, we explicitly state in the Limitations that prospective and live EHR validation are necessary next steps. Importantly, our external validation datasets were drawn from geographically diverse hospitals with heterogenous data structures and patient populations, where the model maintained robust discriminative performance (AUC 0.88-0.96), supporting generalizability beyond a single curated dataset.
We have revised the manuscript to clarify the scope of "real-time" prediction and the current stage of development, noting that prospective validation and EHR integration remain essential next steps. We thank the reviewer again for these constructive comments.

Furthermore, several methodological gaps need to be addressed before firm conclusions can be drawn:
1. Datasets such as EBM-text, PUB-test, and Public-Note would be expected to be in English. What language was used across the internal and external datasets, as well as in the prompts—English or Chinese?
Response: We thank the reviewer for raising this important point regarding language. In our framework, both Chinese and English are used in a deliberately designed bilingual setting with Chinese accounting for the vast majority of the corpus to ensure optimal adaptation to the target

clinical environment.

In the AKICorpus for pre-training, PUB-text (PubMed literature) was entirely in English, EBM-text (guidelines from Up To Date) and EHR-Note (Clinical records from Peking University First Hospital [PKUFH]) were in Chinese, reflecting real-world Chinese clinical practice and documentation. Public-Note, originally from MIMIC-IV and eICU (English), was carefully translated into Chinese to maintain consistency with our Chinese clinical context and prompt format. All fine-tuning and validation datasets (AKILM-FT, internal and external validation) were from Chinese hospitals with Chinese clinical notes. Accordingly, all model prompts were constructed in Chinese.

This bilingual pre-training strategy intentionally incorporates large-scale English medical literature for broad biomedical knowledge, while using Chinese EHR-Note and fine-tuning data to adapt the model to Chinese clinical documentation conventions. The base model, Qwen2.5-7B, is a widely-used bilingual (Chinese-English) architecture that supports this cross-lingual approach.

We have added a dedicated paragraph in the Supplementary Method and Supplementary Table 1 to explicitly clarify the language composition of each data source and the prompting language. The revised text reads as follows:

Supplementary Method Section: *The AKICorpus used for continued pre-training was constructed from both English- and Chinese-language sources. Specifically, PUB-text was entirely in English, whereas EBM-text and EHR-Note were in Chinese. Public-Note, originally derived from the MIMIC-IV and eICU English clinical notes, was translated into Chinese during corpus construction to ensure consistency with the Chinese clinical context and the Chinese prompt format. The AKILM-FT dataset, internal validation dataset, and all external validation datasets were based on Chinese EHRs. All model prompts were formulated in Chinese to ensure semantic alignment with clinical inputs.* (**Page 4, line 65-71**)

Supplementary Table 1. Detailed statistics of datasets for development and evaluation of AKILM.

Dataset	Usage	Description	Size	Language composition
AKICorpus				
EHR-Note	Pre-training	Real-world EHR dataset from AKI	74,891,547 tokens	Chinese
Public-Note	Pre-training	Clinical records from MIMIC-IV and eICU	4,178,424,707 tokens	Chinese
EBM-text	Pre-training	Diverse collection evidence-based medicine guidelines for AKI	917,755 tokens	Chinese
PUB-text	Pre-training	High quality literature on kidney from PubMed	4,813,576 tokens	English
AKILM-FT				
Fine-tuning		Medical record with 6-hour dynamic data from Peking university first hospital	47,750 records	Chinese
Alignment		Medical record with clinician annotated AKI attribution from Peking university first hospital	600 records	Chinese
Evaluation				
Internal validation dataset		In-distribution data from Peking university first hospital	17,074 records	Chinese
External validation dataset		Out-of-distribution data from one tertiary academic hospitals and two local hospitals	75,813 records	Chinese

2. The comparison against generalist LLMs remains methodologically imbalanced since these models were not fine-tuned for structured EHR prediction, whereas AKI-PM was extensively trained on domain-specific clinical data. The reported extreme failure of generalist LLMs (specificity 0.00-0.25 in the main paper) likely reflects mismatched task design, not true inferiority.

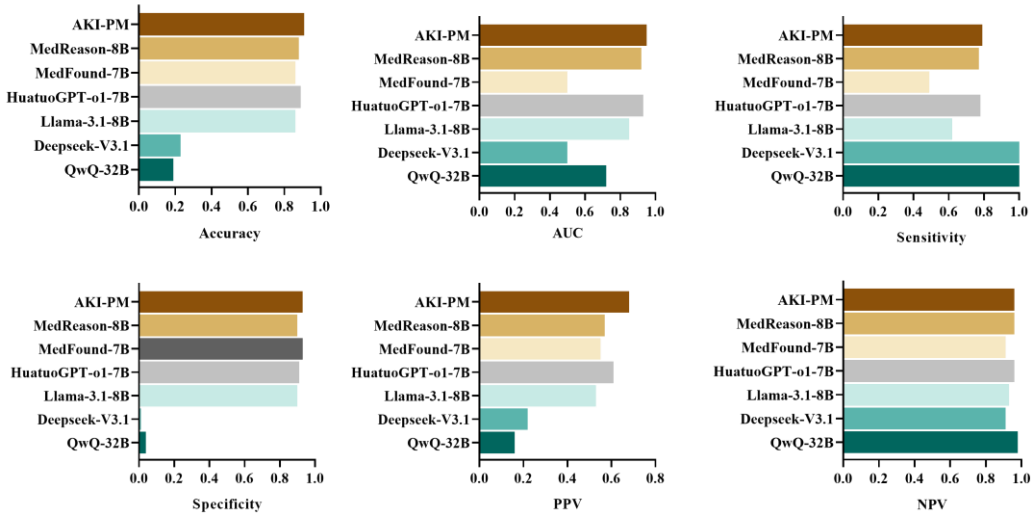
Response: We thank the reviewer for this methodologically important comment. We agree that the original comparison was not fully balanced, as zero-shot evaluation is likely to underestimate the performance of general-purpose LLMs on a structured EHR prediction task. To address this concern, we additionally fine-tuned four baseline models (MedFound-7B, MedReason-8B, HuatuoGPT-o1-7B, and Llama-3.1-8B) using the same task formulation and training split as AKI-PM. Larger models (Deepseek-V3.1 and QwQ-32B) were not fine-tuned due to computational constraints and are retained only as zero-shot reference. Fine-tuning improved the performance of all applicable baseline models, confirming the reviewer's point that the very poor zero-shot specificity partly reflected task mismatch. Importantly, after this training-matched comparison, AKI-PM remained consistently superior across the main discrimination metrics, supporting the conclusion that its advantage arises from kidney-specific continued pre-training and adaptation to structured longitudinal EHR data, not merely from benchmark design.

We have updated **Figure 2** with the fine-tuned results and revised the **Methods and Results sections** accordingly, with clear annotation distinguishing zero-shot from fine-tuned conditions.

Methods Section: *To ensure a more balanced comparison, we additionally fine-tuned four open-access baseline LLMs (MedFound-7B, MedReason-8B, HuatuoGPT-o1-7B, and Llama-3.1-8B) using the same AKI prediction task formulation and training split as AKI-PM. Larger models (Deepseek-V3.1 and QwQ-32B) were not fine-tuned because of computational constraints and were retained only for zero-shot reference comparison.* **(Page 18, line 382-387)**

Results Section: *In zero-shot evaluation (Supplementary Fig. 3), four models (MedReason-8B, QwQ-32B, Deepseek-V3.1, and HuatuoGPT-o1-7B) exhibited marked performance imbalances, whereas AKI-PM outperformed the remaining three. To enable a more balanced comparison, we additionally fine-tuned four baseline models (MedFound-7B, MedReason-8B, HuatuoGPT-o1-7B, and Llama-3.1-8B) using the same task formulation and training split as AKI-PM. The larger models (Deepseek-V3.1 and QwQ-32B) were retained as zero-shot references only. Fine-tuning improved all applicable baselines, particularly in accuracy, specificity and PPV. Nevertheless, AKI-PM remained the best-performing model overall, with the strongest PPV (Fig. 2).* **(Page 7, line 128-136)**

Figure 2. Comparative performance of AKI-PM and other LLMs.



Predictive performance comparison for dynamic AKI prediction within 24 hours among three general open-access LLMs (Llama-3.1-8B, QwQ-32B, and Deepseek-V3.1), three medical open-access LLMs (HuatuoGPT-o1-7B, MedReason-8B, and MedFound-7B), and our AKI-PM in internal validation dataset (n=17,074). The MedFound-7B, MedReason-8B, HuatuoGPT-o1-7B, and Llama-3.1-8B were fine-tuned using the same AKI prediction task formulation and training split as AKI-PM. Deepseek-V3.1 and QwQ-32B were not fine-tuned due to computational constraints and are reported only as zero-shot reference comparisons. The model performance was assessed using AUC, accuracy, sensitivity, specificity, PPV, and NPV. AKI-PM: acute kidney injury - prediction model; LLM: large language model; AUC: area under curve; PPV: positive predictive value; NPV: negative predictive value.

3. Regarding AKI-RAM, does the model restrict its attention to the examples included in the prompt? For instance, when asked to explain intervenable risk factors, does it limit its response to risk elements that clinicians could realistically modify?

Response: We thank the reviewer for raising this important question. AKI-RAM does not restrict its output to the specific examples provided in the prompt. Instead, the prompt offers representative clinical anchors that help the model distinguish modifiable from non-modifiable factors, without constraining its reasoning scope. Specifically, it includes illustrative examples of modifiable factors (e.g., infection, hypovolemia, progressive anemia, and nephrotoxic medications) and non-modifiable factors (e.g., age, gender, baseline kidney function, pre-existing comorbidities, and surgical history). These examples guide clinical interpretation but do not serve as a predefined whitelist.

To evaluate this capability, we conducted a human validation on 200 representative cases. AKI-RAM identified risk factors and classified them as modifiable or non-modifiable per clinical consensus, consistently identifying modifiable factors beyond the prompt examples, such as acute heart failure, thrombocytopenia, and severe electrolyte disturbances, and classifying them appropriately (as shown in the examples bellow). These findings indicate that AKI-RAM performs generalized clinical reasoning based on patient-specific context, rather than relying on simple

pattern matching to prompt examples.

Example 1: The AKI-RAM Outputs the hyponatremia and thrombocytopenia in modifiable risk factors

[Non-modifiable Risk Factors]
- History of surgery (276–282 hours) => Surgical trauma and stress may further increase renal burden.
[Modifiable Risk Factors]
- Use of nephrotoxic drugs (aminoglycosides, sulfonamides, vancomycin, tetracyclines) => Concomitant nephrotoxic agents may result in additive renal injury. - Serum sodium (150 h 136.61 → 234 h 129.52 → 282 h 123.92 → 288 h 123.3) => Progressive hyponatremia, the volume status should be closely monitored. - Neutrophil percentage (150 h 99.3% → 282 h 98.9%), high-sensitivity C-reactive protein (150 h 10.24 → 282 h 186.94), PCT (PCT: 150 h 0.33 → 276 h 6.02) => Severe infection need be treated with antimicrobial therapy. - Hemoglobin (150 h 87 → 180 h 90 → 210 h 78 → 234 h 76 → 282 h 65) => Persistent severe anemia may influence the oxygen delivery to kidney. - Thrombocytopenia (150 h 179 → 180 h 119 → 210 h 57 → 234 h 26 → 258 h 18) => Profound thrombocytopenia, hemorrhage prevention is essential, and platelet transfusion should be considered.
[Integrated Clinical Recommendations]
- Immediately discontinue non-essential nephrotoxic drugs (e.g., aminoglycosides, vancomycin) and preferentially switch to renally safer alternatives. - For severe sepsis, escalate antimicrobial coverage, dynamically monitor inflammatory markers, and initiate fluid resuscitation when necessary. - Correct hyponatremia and anemia, maintain effective circulating blood volume, and avoid exacerbation of pre-renal AKI. - Assess coagulation function, remain vigilant for bleeding risk, and transfuse platelets when appropriate.

Example 2: The AKI-RAM Outputs the acute heart failure in modifiable risk factors

[Non-modifiable Risk Factors]
- Age 81 years => advanced age is an independent risk factor for AKI.
[Modifiable Risk Factors]
- Vancomycin, anticancer agents, proton pump inhibitors, NSAIDs (12–18 hours), diuretics (66–72 hours) => Concomitant use of nephrotoxic drugs may lead to renal tubular necrosis or interstitial nephritis. - BUN/creatinine ratio (24–30 hours: 8.27/87.8 → 42–48 hours: 9.41/81.8) => The ratio gradually increases and a value >20 suggests insufficient effective circulating volume. - Neutrophil percentage (24–30 hours: 90.2% → 42–48 hours: 84.2%), high-sensitivity C-reactive protein (24–30 hours: 24.86 mg/L → 42–48 hours: 174.2 mg/L) => Persistent inflammatory state can impair renal perfusion. - BNP (24–30 hours: 254 pg/mL → 42–48 hours: 729 pg/mL → 114–120 hours: 1102 pg/mL) => Progressive deterioration of cardiac function can cause kidney injury.
[Integrated Clinical Recommendations]
- Immediately discontinue non-essential nephrotoxic medications (e.g., vancomycin, NSAIDs) and switch to safer alternatives with better renal tolerability. - Strengthen antimicrobial therapy, and adjust antibiotics based on microbiological culture results. - Implement precision fluid management, with dynamic monitoring of serum creatinine, urine output, and electrolytes.

4. Regarding the df2text pipeline (Supplementary Method Section 3.1 and Supplementary Figure 5), several critical details remain unaddressed:

a. df2text does not address time-interval weighting or encoding of irregular sampling beyond linear text ordering. How does the model account for irregular sampling intervals beyond linear ordering?

Response: We appreciate the reviewer's careful evaluation of *df2text* pipeline. In our study, the dynamic EHR data used for model development were aggregated into fixed 6-hour intervals before being processed by *df2text*. This resulted in a uniformly sampled temporal sequence, and we relied on the linear text order to encode time progression without introducing additional encoding or weighting for irregular time gaps. The *df2text* module converted these pre-aligned 6-hour data blocks into a semantically structured narrative format that LLMs can process naturally.

For missing values within a given 6-hour window, we chose not to perform explicit statistical imputation to avoid introducing artificial bias. Instead, *df2text* converts only the observations that were actually recorded. This design allows the LLM to handle missingness and data sparsity through contextual reasoning, while preserving the original characteristics of real-world clinical documentation. We have clarified this uniform temporal structure and our handling of sparsity in the revised Supplementary Methods.

Supplementary Methods Section: *All longitudinal EHR data were aggregated into fixed 6-hour intervals before df2text conversion, yielding a uniformly sampled time series. The linear ordering of the generated text therefore encoded time progression without requiring additional time-interval weighting. For missing values, df2text verbalized only the available observations and did not perform explicit statistical imputation, allowing the LLM to interpret data sparsity through contextual reasoning.* (**Page 7, line 183-187**)

b. The truncation to 16 rows (Supplementary Method Section 3.3) may introduce information loss. What is the effect of truncating to 16 rows on prediction accuracy and how many patients experience significant loss of historical data?

Response: We appreciate the reviewer's thoughtful concern. In our framework, the truncation to 16 rows was designed to balance information retention with efficient LLM input length. Each row represents a 6-hour time window containing both static baseline variables and the most recent dynamic measurements. Retaining the 16 most recent rows therefore preserves up to 96 hours of longitudinal data, while baseline profile remains fully available at each prediction time point. We considered this window clinically reasonable, as recent physiological trajectories are generally more informative for AKI prediction than more remote history.

We also quantified the extent of truncation in the study cohort: 51.73% of patients had longitudinal records exceeding 16 rows and thus underwent truncation. For these patients, because full baseline information and the most recent 96 hours of dynamic data were preserved, the discarded portion consisted primarily of more remote historical observations. To assess the impact of this truncation, we conducted sensitivity analyses using truncation lengths of 16, 32, and 64 rows. Predictive performance remained materially unchanged (accuracy, 0.91-0.92; AUC, 0.95-0.96), indicating that truncation to 16 rows did not lead to meaningful information loss. We have clarified this rationale and added the corresponding sensitivity analyses to the revised Supplementary Method and Supplementary Table 10.

Supplementary Methods Section: *Each compressed table was truncated to a maximum of 16 rows before being converted into text input for the model, which balanced clinical practicality and the context-length limitation with the preservation of relevant temporal information.* (**Page 8, line 254-256**)

Supplementary Table 10. Performance of AKI-PM with different truncation lengths of rows.

Rows	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
16	0.95	0.91	0.79	0.93	0.68	0.96
32	0.96	0.92	0.83	0.93	0.69	0.97
64	0.96	0.92	0.83	0.94	0.70	0.97

AKI-PM, acute kidney injury - prediction model; AUC, area under curve; PPV, positive predictive value; NPV, negative predictive value.

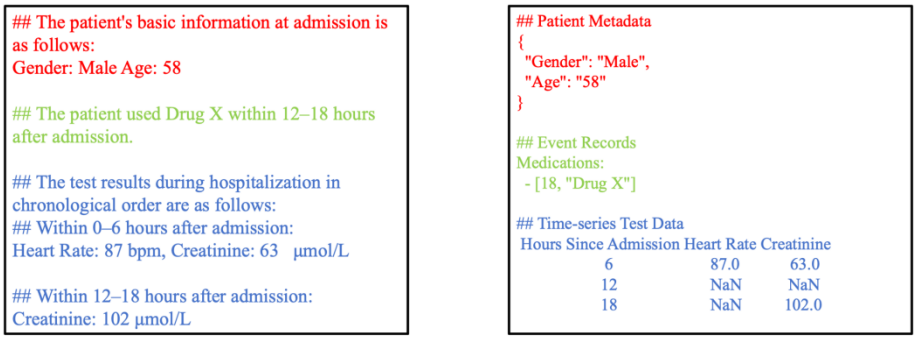
c. The supplement demonstrates a 0.94 to 0.81 AUC drop (Accuracy 0.90 vs 0.62, Specificity 0.92 vs 0.58) when converting df2text to df2string (Supplementary Table 8), yet does not explain why string formatting is so detrimental.

Response: We thank the reviewer for highlighting this performance difference. The main reason is that *df2text* converts structured EHR data into semantically coherent, context-rich natural language.

This representation helps clarify the relationships among variables, values, and time points more explicitly, thereby preserving temporal logic and supporting contextual clinical reasoning. In contrast, *df2string* is a linear concatenation of tabular entries, in which variables and values are appended sequentially with limited syntactic or semantic structure. As a result, the input tends to be more fragmented and less interpretable, making it more difficult for the model to identify meaningful patterns and infer cross-variable relationships. This difference is particularly important for LLMs, which are pre-trained predominantly on natural language corpora and are therefore better aligned with structured textual narratives than with compressed symbolic strings. The substantial decline in AUC, accuracy, and specificity with *df2string* therefore suggests as a representational mismatch rather than loss of clinical information. We have added this rationale to the Discussion section in the revised manuscript and included an example of *df2text* and *df2string* in Supplementary Figure 7.

Discussion Section: *Moreover, compared with simple string concatenation, our structured textual input representation better aligns longitudinal EHR data with LLMs, preserving temporal and cross-variable relationships and thus improving predictive performance.* (**Page 13-14, line 280-283**)

Supplementary Figure 7. An example of *df2text* and *df2string*



5. The pretraining corpus (AKICorpus) includes institutional EHR-Note data drawn from the same center used for fine-tuning and internal validation, raising the possibility of subtle leakage of institutional patterns, coding habits, or lexical conventions.

Response: We thank the reviewer for raising this important concern. We apologize for a typographical error in the Supplementary Material that may have caused confusion. The EHR-Note dataset used for pretraining and fine-tuning was collected from Peking University First Hospital from 2018 to 2019, whereas the internal validation set was collected strictly from 2020. Therefore, there was no patient-level overlap between the two datasets.

We acknowledge that using data from the same institution for pretraining and internal validation could theoretically allow the model to learn institution-specific patterns. However, the robust performance observed across four geographically distinct external validation hospitals with different EHR systems and clinical documentation habits indicates that the model's predictive ability is not driven by center-specific bias.

We have corrected the corresponding dates in the revised manuscript and Supplementary Materials for clarity.

Method Section: *a proprietary dataset of 47,750 admissions from Peking University First Hospital (PKUFH) in China (EHR-Note).* (**Page 16, line 339-341**)
Supplementary Method Section: *EHR-Note is a proprietary, large-scale, real-world dataset sourced from Peking University First Hospital in China, with 47,750 records from January 1, 2018, to December 31, 2019.* (**Page 3, line 22-23**)

6. The deduplication thresholds (α=0.7, β=0.88) may allow retention of semantically similar clinical trajectories. Whether were near-duplicate cases in the EHR corpus inspected or filtered?

Response: We thank the reviewer for this important comment. For the first-stage exact-match deduplication, according to the default configurations of Data-Juicer[1], a widely used data processing framework for LLMs, the Jaccard_threshold for MinHash-based document deduplication is universally set to 0.7. This specific threshold effectively captures literal near-duplicates caused by repetitive EHR templates or minor formatting variations, without aggressively discarding records that share common medical vocabulary but represent different patient events.

For the second-stage semantic-level deduplication, we adopted a threshold of β = 0.88 based on empirical validation. To further assess its adequacy, we conducted a manual review of 100 randomly sampled case pairs with similarity scores near the retention threshold (β = 0.80-0.88). These pairs were examined by a clinical expert, who found that although some cases shared common template-like clinical expressions, their core clinical trajectories, temporal evolution, and key events were distinct, indicating that they were not duplicate patient records. For example, here are two text statements with β = 0.88:

Text 1: Below is information on surgeries, examinations, or medications performed for this patient during hospitalization: Examination: Hematocrit; Blood; Result: 33.3% …

Text 2: Below is information on surgeries, examinations, or medications performed for this patient during hospitalization: Surgery: Placement of intercostal catheter for drainage; Surgery date/time: 2015/2/15 …

Together, these findings suggest that our original thresholds (α = 0.7, β = 0.88) were sufficient to remove true duplicates while preserving clinically meaningful variation. We have clarified this point in the revised Supplementary Method.

Supplementary Methods Section:

-Exact-match deduplication: According to the default configurations of Data-Juicer, MinHash signatures of text blocks were computed, and duplicates were removed using a Jaccard similarity threshold of α = 0.74.

-Semantic-level deduplication: Sentence-BERT5 was used to generate text embeddings, and semantically redundant content was removed using a cosine similarity threshold of β = 0.88. The value of β was determined through empirical validation and corroborated by clinical assessment from physicians. (**Page 8, line 232-238**)

Reference:

[1] Chen, D. et al. Data-Juicer: A One-Stop Data Processing System for Large Language Models. arXiv:2309.02033 (2023).

7. The largest component is drawn from MIMIC-IV and eICU—nearly entirely ICU-heavy datasets, yet the model systematically performs worse in ICU populations across all centers (Supplementary

Tables 4–7). How to explain this asymmetry and potential disparities?

Response: We thank the reviewer for this insightful comment. We agree that the ICU-heavy composition of MIMIC-IV and eICU might seem inconsistent with the relatively lower performance of our model in ICU subgroups. We would like to offer the following clarifications.

First, MIMIC-IV and eICU were used for domain pre-training, to expose the model to kidney-related and critical care clinical language and knowledge, rather than for task-specific decision boundary learning. In our experiments, task-specific fine-tuning was the main determinant of predictive performance: fine-tuning alone substantially improved model performance compared with pre-training alone (Accuracy, 0.89 vs. 0.48; AUC, 0.93 vs. 0.60). Importantly, the fine-tuning cohort consisted predominantly of general ward (non-ICU) patients. As a result, the model's decision boundaries are more aligned with non-ICU populations, which likely contributes to the observed performance gap in ICU subgroups.

Second, AKI prediction tends to be more challenging in ICU patients, whose clinical course is often characterized by rapidly evolving and multifactorial critical illness, with complex interactions among hemodynamic instability, sepsis, organ support, and medications. Similar performance attenuation in ICU subgroups has been reported in prior hospital-wide AKI prediction studies [1], suggesting that this pattern reflects the intrinsic complexity of ICU patient populations rather than a model-specific bias.

We have added this interpretation to the revised Discussion and acknowledged lower performance in ICU patients as a limitation and a priority for future ICU-specific optimization.

Discussion Section: *The performance of AKI-PM is less robust in ICU settings, likely reflecting the greater clinical complexity of critical illness and the fact that domain pre-training alone does not fully capture case mix. Future work should therefore prioritize context-aware adaptation (e.g., ICU-specific fine-tuning) to improve accuracy and reliability in critically ill patients.* **(Page 14, line 295-299)**

Reference:

[1] Zhang Y, et al. Development and validation of a real-time prediction model for acute kidney injury in hospitalized patients. Nat Commun. 2025;16(1):68. Published 2025 Jan 2. doi:10.1038/s41467-024-55629-5

8. The manuscript highlights few-shot improvement for external centers, but the supplement does not describe: How those 1,000 samples were selected or distributed across AKI vs non-AKI? Whether prompting, calibration, or underlying weights were modified? Whether smaller adaptation sets (e.g., 10–200 samples) were tested?

Response: We thank the reviewer for highlighting this point. For each external center, the 1,000 adaptation samples were randomly selected without class rebalancing, so that the AKI/non-AKI distribution reflected the original prevalence in that center. For few-shot adaptation, we updated the model weights via LoRA-based fine-tuning in LLaMA-Factory [1] using the selected 1,000 samples, without additional prompting or calibration. Detailed hyperparameters and training configurations have been added to Supplementary Table 4.

As suggested, we tested adaptation set sizes of 10, 100, 200, 500, and 1,000 samples per external center. Model performance improved progressively with sample size, with meaningful gains already observed at 100 samples, and near-plateau performance reached at approximately 1,000 samples.

These results have been added to Supplementary Table 11, and we have clarified the few-shot adaptation procedure in the Method.

Reference:

[1] Zheng, Y. *et al.* LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. *ArXiv* abs/2403.13372, (2024).

Method Section: *For few-shot adaptation at each external center, we randomly sampled 1,000 cases without class rebalancing to preserve the original AKI/non-AKI prevalence, and performed LoRA-based37 fine-tuning in LLaMA-Factory[35] to update model weights without additional prompting or post hoc calibration. Detailed hyperparameters and training configurations have been added Supplementary Table 4. To assess sample efficiency, a gradient experiment was conducted using adaptation sets of 10, 100, 200, 500, and 1,000 samples (Supplementary Table 11).* (**Page 19-20, line 413-419**)

Supplementary Table 4. The parameters of configuration of few-shot.

Parameters	**Hospital**		
	Miyun	Taiyuan	Sichuan
cutoff_len	30000	30000	30000
preprocessing_num_workers	8	8	8
per_device_train_batch_size	1	1	1
gradient_accumulation_steps	2	2	2
learning_rate	2.0e-4	1.0e-4	2.0e-4
num_train_epochs	8.0	5.0	5.0
lr_scheduler_type	cosine	cosine	cosine
warmup_ratio	0.1	0.1	0.1
precision	bf16	bf16	bf16

Supplementary Table 11. The gradient experiment for few-shot across different samples

Samples	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
Taiyuan cohort						
zero-shot	0.90	0.87	0.57	0.93	0.60	0.92
10	0.91	0.88	0.64	0.92	0.61	0.93
100	0.92	0.90	0.60	0.95	0.70	0.93
200	0.93	0.90	0.66	0.95	0.69	0.94
500	0.94	0.90	0.70	0.94	0.69	0.94
1000	0.94	0.91	0.67	0.95	0.73	0.94
Miyun cohort						
zero-shot	0.91	0.88	0.59	0.93	0.62	0.92
10	0.91	0.88	0.55	0.95	0.65	0.92
100	0.92	0.89	0.62	0.94	0.65	0.93
200	0.92	0.89	0.56	0.95	0.67	0.92

500	0.93	0.90	0.62	0.95	0.70	0.93
1000	0.92	0.90	0.67	0.94	0.69	0.94
Sichuan cohort						
zero-shot	0.88	0.82	0.75	0.84	0.47	0.95
10	0.89	0.84	0.73	0.87	0.51	0.94
100	0.93	0.89	0.75	0.92	0.63	0.95
200	0.95	0.91	0.70	0.95	0.73	0.94
500	0.96	0.92	0.78	0.95	0.76	0.96
1000	0.96	0.92	0.77	0.95	0.74	0.96

AUC, area under curve; PPV, positive predictive value; NPV, negative predictive value.

9. The rejected responses were generated by DeepSeek-R1. Whether rejected responses might systematically bias the model toward particular reasoning patterns rather than toward universally valid clinical logic?

Response: We thank the reviewer for raising this insightful question. Large language models are not inherently constrained to a fixed reasoning pattern; their outputs are driven by the prompt and the clinical context provided. In our study, the rejected responses were generated by DeepSeek-R1 using the same AKI-RAM prompt (Supplementary Methods 2.2). This prompt does not impose a specific reasoning template or constrain the internal logic of the model; it provides general clinical guidance with valid reasoning examples (e.g., nephrotoxic drug use, hypovolemia, anemia, and infection). Therefore, the prompt was not designed to bias DeepSeek-R1 toward any "particular reasoning patterns", and the rejected responses were generated based on general clinical reasoning rather than artificial structural constraints.

To directly evaluate potential systematic bias, we reviewed 600 rejected responses produced by DeepSeek-R1 and subsequently corrected by clinical experts to create the corresponding "corrected chosen responses". We found that the errors were highly diverse and did not concentrate in a single reasoning pattern. Specifically, five major categories of errors were observed:

(1) Erroneously identifying a non-risk factor as relevant, e.g., classifying an age of 39 as an AKI risk factor.

(2) Confusing modifiable with non-modifiable factors, e.g., miscategorizing nephrotoxic medication use as a non-modifiable risk factor.

(3) Deriving an incorrect conclusion from given evidence, e.g., interpreting a minor, non-significant change in neutrophil percentage from 74.3% to 75.2% over 24 hours as definitive evidence of infection or inflammation.

(4) Inaccurate clinical definitions, e.g., classifying a serum potassium level of 3.88 mmol/L as hypokalemia.

(5) Missed key risk factors like infection.

This heterogeneous error distribution suggests the model errors reflect genuine, multifaceted deficiencies in clinical knowledge and logical reasoning, rather than a systematic bias toward particular reasoning patterns. The correction process is designed to address exactly these types of errors.

10. Supplementary Method Section 2 describes deterministic prompt templates, but the inference

section (Section 6) notes that invalid outputs require repeated attempts. Please report the failure rate across the full validation set and consider implications for deployment stability.

Response: We thank the reviewer for raising this critical point regarding real-world deployment stability. In response, we have systematically quantified the parsing failure rate across the full validation set. On the initial attempt, 21.2% of the samples required regeneration due to minor formatting anomalies (e.g., malformed JSON structures or extraneous text). These initial failures reflect structural formatting deviations inherent to auto-regressive generative models, rather than errors in the model's clinical reasoning or predictive capability. We allow up to three regeneration attempts, which effectively ensures valid outputs for downstream parsing. Regarding the implications for deployment stability, this retry mechanism introduces a modest increase in computational latency but does not compromise the overall robustness or reliability of the pipeline.

11. Supplementary Method Section 6 indicates that inference may require up to three passes before a valid response is generated. Whether this affects real-time prediction feasibility and how output validation is handled in live settings?

Response: We thank the reviewer for raising this important point. Our design allows up to three generation attempts for AKI-RAM to ensure a valid and clinically relevant response. In evaluation, nearly 80% of cases are resolved within a single pass, with an average inference latency of 4.13 s. Even when accounting for the maximum allowance of three attempts, the overall average latency is 25.08 s, which we believe is acceptable for clinical workflows that require actionable risk information rather than instantaneous alerts.

In real-world deployment, our system operates in a two-stage manner. First, AKI-PM identifies patients at high risk of AKI within the next 24 hours and triggers a pop-up alert. The average latency for this high-risk identification is only 0.23 s, which is effectively real-time for clinicians. For those alerted high-risk patients, AKI-RAM then generates relevant, potentially modifiable risk factors and clinical intervention recommendations. While the additional response time of 25.08 s is longer than the alert timing, it is still acceptable in practice because it supports decision-making and care planning rather than immediate event-by-event prediction. In addition, the output validation in the live setting is handled through an automated post-generation check against a predefined structured-response schema. If an output was incomplete or unparsable, the model was prompted again, for up to three attempts, and only responses that passed validation were accepted for downstream prediction.

12. Supplementary Figure 3 provides a sample AKI-RAM output, but interpretation concerns remain: The attribution schema is heavily prompt-driven rather than model-internal. It is not possible to determine whether AKI-RAM identifies genuine causal mechanisms or merely reorders evidence extracted from df2text. Evaluation of AKI-RAM also lacks inter-rater agreement and external raters to mitigate institutional bias.

Response: We sincerely thank the reviewer for this rigorous evaluation of the AKI-RAM model. We fully agree that distinguishing genuine clinical reasoning from prompt-induced text organization is essential.

LLMs are inherently prompt-driven [1-2]. While our prompt template standardizes the output structure for clinical readability, it does not predetermine the clinical conclusion. The prompt serves

as a contextual trigger that activates the model's internal medical knowledge. The prompt serves to elicit structured reasoning and attribution from the model, but does not itself encode the specific clinical conclusions. In AKI-RAM, the identification of risk factors is driven entirely by its internal weights, which were rigorously optimized during the expert-guided preference learning phase. Consequently, the model does not merely extract or reorder the input text; it actively synthesizes temporal trajectories across modalities. For example, it can infer physiological causality by linking specific diuretic exposure to subsequent elevations in SCr and BUN, a deductive process that extends well beyond surface-level text rearrangement. The prompt functions solely as a task specification interface, whereas the substantive clinical attribution is determined entirely by the model's learned representations and reasoning capacity.

To address the concern regarding institutional bias, we expanded our clinical assessment panel by recruiting three additional independent nephrologists, each with more than 5 years of clinical experience, from two external local hospitals. We then rigorously quantified the inter-rater reliability among six physicians across the eight dimensions. The evaluation demonstrated moderate to good agreement with the average-measures ICC ranging from 0.680 to 0.803, confirming that the causal reasoning generated by AKI-RAM aligns with broadly accepted nephrology principles rather than institution-specific heuristics. We have updated the revised Methods and Results sections to include the details of this expanded multi-center expert panel and the corresponding agreement metrics.

Methods Section: *Evaluation was performed by six experienced nephrologists (each with >5 years of clinical experience) using a 0-5 Likert scale on 200 cases drawn from both the internal and external validation datasets. The panel comprised three experts from an academic hospital and three from two independent local hospitals. Inter-rater reliability across six assessors was quantified using intraclass correlation coefficients (ICCs) across eight dimensions. ICCs were calculated in SPSS using a two-way random-effects model with absolute agreement and average measures.* (**Page 19, line 404-410**)

Results Section: *To evaluate the real-world clinical utility of AKI-RAM, six nephrologists assessed its performance using a 0-5 Likert scale across eight clinical dimensions (Fig. 4). In internal validation dataset, AKI-RAM achieved scores of 4.27 for medical case comprehension, 4.52 for guideline adherence, and 4.18 for clinical reasoning. For clinical decision support, it scored 4.55 for relevance of risk factors and 4.74 for acceptability of clinical recommendations. For risk control, it achieved scores of 4.87, 4.88 and 4.86 for mitigating the risks associated with unfaithful content, bias and unfairness, and possibility of harm, respectively. Performance remained consistent across all three external validation cohorts. The inter-rater reliability among six nephrologists across eight dimensions (average-measures ICC) ranged from 0.680 to 0.803, indicating moderate to good agreement (Supplementary Table 12).* (**Page 8-9, line 170-180**)

Supplementary Table 12. Inter-rater reliability across the eight dimensions

	Average measures ICC	95% CI
Medical case comprehension	0.785	0.736-0.828
Guideline adherence	0.803	0.757-0.842
Clinical reasoning	0.758	0.702-0.806
Risk factor relevance	0.751	0.694-0.801
Recommendation acceptability	0.695	0.625-0.756
Unfaithful content	0.705	0.637-0.764
Bias and fairness	0.680	0.607-0.744
Possibility of harm	0.727	0.665-0.782

ICC: Intraclass Correlation Coefficient; 95% CI: 95% Confidence Interval

Reference:

[1] Brown, T. et al. Language models are few-shot learners. Advances in neural information processing systems (NeurIPS). arXiv:2005.14165 (2020).

[2] Liu, P. et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv:2107.13586 (2023).

Reviewer #3 (Remarks to the Author):

This is a nice attempt to develop a model for AKI prediction across all hospitalized patients. The model works fine, and most importantly, it provides modifiable factors for potential intervention. Though not prospectively validated in the real world, it is still worth being published to pave the way for future validation. I am not an expert in LLM, but I have several comments to further improve the manuscript based on my expertise in clinical studies.

Response: We sincerely thank the reviewer for their encouraging assessment and for recognizing the clinical value of our work, particularly the identification of modifiable risk factors. We fully agree that prospective validation is the critical next step, and we hope this study provides a robust foundation for that effort. We appreciate the reviewer's expert insights that have strengthened the manuscript and we address each comment below.

1. The previous studies were both conducted in enriched populations, so this study is a major step forward. This should be mentioned in the Introduction.

Response: We thank the reviewer for this constructive suggestion. We agree that the distinction between enriched and unselected patient populations is a critical methodological consideration. We have revised the Introduction to explicitly highlight this gap, emphasizing that our model applies to all hospitalized patients regardless of admission type or pre-existing risk stratification.

Introduction Section: *However, most previously AKI prediction models were developed in enriched populations with inherently higher baseline risk (e.g., ICU or cardiac surgery patients), limiting their generalizability to the broader inpatient population. Among the few models developed for general hospitalized patients, machine learning models, including our own prior work[9], have achieved good discrimination[area under curve (AUC) values 0.88-0.96] but are constrained by low positive predictive values (PPV: 0.06-0.30)[9-14], leading to frequent false alarms; and by the lack of capability to translate feature importance into actionable interventions. These limitations highlight the need for novel approaches that can improve predictive precision while providing*

interpretable, clinically actionable outputs. (**Page 4, line 72-80**)

2. Line 331-332，how the authors define "nephrologist";

Response: We thank the reviewer for this helpful comment. We agree that “nephrologist” should be explicitly defined for clarity and reproducibility. In the revised manuscript, we now specify that a nephrologist refers to a licensed physician practicing in the Department of Nephrology with more than 5 years of independent clinical experience. All physicians involved in the expert evaluation met this criterion. We have updated the Supplementary Methods to clarify the definition.

Supplementary Methods Section: *A nephrologist was defined as a licensed physician practicing in the Department of Nephrology with more than 5 years of independent clinical experience. For each case, two nephrologists conducted a comprehensive review of medical records to identify risk factors associated with AKI, categorizing each factor as modifiable or non-modifiable. Personalized clinical recommendations were subsequently formulated based specifically on the modifiable risk factors. In case of disagreement, a third nephrologist with over 10 years of experience was consulted.* (**Page 3-4, line 44-50**)

3. Ethical approval should be obtained in all the participating sites, including those for external validation.

Response: We thank the reviewer for this important point. This study is a secondary analysis of a previously established multi-center database, for which ethical approval had already been obtained from all participating sites at the time of database construction, as reported in our prior work [1].

For the present secondary analysis, we obtained renewed ethical approval from our primary center (PKUFH, approval number 2025[1310-001]). In accordance with the reviewer’s recommendation, we have now additionally obtained ethical approval from all external centers (Sichuan, approval number 2026[208]; Miyun, approval number 2026[005-001]; Taiyuan, approval number 2026008). The Methods section has been updated to list all approval numbers. Informed consent was waived in accordance with institutional guidelines for retrospective, de-identified data analyses.

Methods Section: *The protocol for the present analysis was approved by the Research Ethics Committee of Peking University First Hospital (approval number 2025-1310), Sichuan Provincial People’s Hospital (approval number 2026[208]), Beijing Miyun District Hospital (approval number 2026[005-001]), and Taiyuan Central Hospital (approval number 2026008). The requirement for patient informed consent was waived due to the retrospective nature of the study, which only involved the collection of existing and de-identified medical data.* (**Page 16, line 328-333**)

Reference:

[1] Zhang Y, et al. Development and validation of a real-time prediction model for acute kidney injury in hospitalized patients. Nat Commun. 2025;16(1):68. Published 2025 Jan 2. doi:10.1038/s41467-024-55629-5

4. How the authors define modifiable factors should be detailed;

Response: We thank the reviewer for this important comment. In our study, a factor was defined as "modifiable" if it can be altered, managed, or reversed through timely physician intervention during hospitalization. To operationalize this definition for the Large Language Model (AKI-RAM), we

employed in-context learning by providing explicit positive examples (infection, hypovolemia, progressive anemia, and nephrotoxic medications) and negative examples (age, gender, baseline kidney function, comorbidities, and surgical history) in the prompt. This approach guided the model to distinguish actionable targets from static baseline risks.

We have added this definition to the Supplementary Methods and provided the exact prompts in the Supplementary Information (**Page 5, line 115-177**).

Supplementary Methods Section: *Modifiable risk factors were defined as clinical variables that can be altered, managed, or potentially reversed through timely physician intervention during the course of hospitalization.* (**Page 9, line 274-276**)

5. What is the weight of the two modifiable factors, "insufficient fluid" and "new infection" in the prediction? Can they be quantified? If the weight is too low, the clinical benefits may be just marginal with extra human resource cost.

Response: We thank the reviewer for this thoughtful and practical question. We fully agree that understanding the weight or relative contribution of each modifiable factor is important for clinical decision-making. The LLM-based interpretability model (AKI-RAM) does not assign numerical weights to individual risk factors in the conventional sense, as would a linear regression model. However, to help clinicians assess the potential impact of these factors, the model presents temporal trends of relevant parameters as supporting evidence. For example, for a factor such as "new infection", the model provides the trajectory of inflammatory markers and vital signs, as shown in Supplementary Figure 4. This allows clinicians to better assess the severity of each factor and thus decide whether intervention is likely to be beneficial, avoiding unnecessary effort for low-impact factors. Our future work will build on this interpretability foundation, and deeply explore the causal relationships between identified modifiable risk factors and AKI onset, aiming to quantify the potential change in AKI probability associated with interventions targeting specific factors. We have clarified this design in the manuscript.

Discussion Section: *Finally, our current AKI-RAM offers evidence-based temporal insights for modifiable factors rather than directly quantifying their individual contributions as numerical weights.* (**Page 14, line 301-303**)

6. What does the factor "new-onset infection" mean? or what is it based on? If it is based on medical records, I don't think it is really "modifiable". Once the diagnosis is made by the physician, treatment usually ensues. If it is based on microbiology findings and lab results, please also report its consistency with clinical diagnosis.

Response: We thank the reviewer for this highly insightful comment. We fully agree that distinguishing the basis of "new-onset infection" is clinically important and we are glad to clarify.

In our study, the identification of new-onset infection was derived exclusively from objective, structured clinical data, specifically serial laboratory measurements (white blood cell count, neutrophil count, C-reactive protein) and microbiological culture results, rather than from physician-authored narrative notes. Therefore, the model does not simply flag infections already diagnosed and treated, instead, it helps identify patients who may have developing but not yet clinically overt infection. For such patients, early recognition can prompt timely antimicrobial

therapy or source control, potentially mitigating AKI risk. We have clarified this point in the Supplementary Methods and provided a detailed example in Supplementary Figure 4.

To directly address the reviewer's question regarding the consistency between the model's inference and actual clinical diagnoses, a nephrologist, defined as a licensed physician practicing in the Department of Nephrology with more than 5 years of independent clinical experience, reviewed the same 50 randomly selected cases from the PKUFH cohort (the same subset utilized for the AKI-RAM manual evaluation) and judged the presence of new-onset infection according to the predefined diagnostic criteria [1]: the definition of infection is satisfied by either (3) or both (1) and (2): (1) Symptoms or signs of infection; (2) Objective supportive evidence from imaging and laboratory tests; (3) Positive etiological evidence. The clinician's judgements were then compared with the model's output, yielding a concordance rate of 88%.

Supplementary Method Section: *All risk factors presented by AKI-RAM, especially infection, are systematically derived from objective, routinely collected clinical data. These encompass quantifiable metrics such as serial laboratory test results, vital signs, medication administration records, and other structured electronic health record entries. This principled reliance on objective inputs ensures that the model's outputs are directly actionable for clinicians.* (**Page 5, line 109-113**)

Reference:

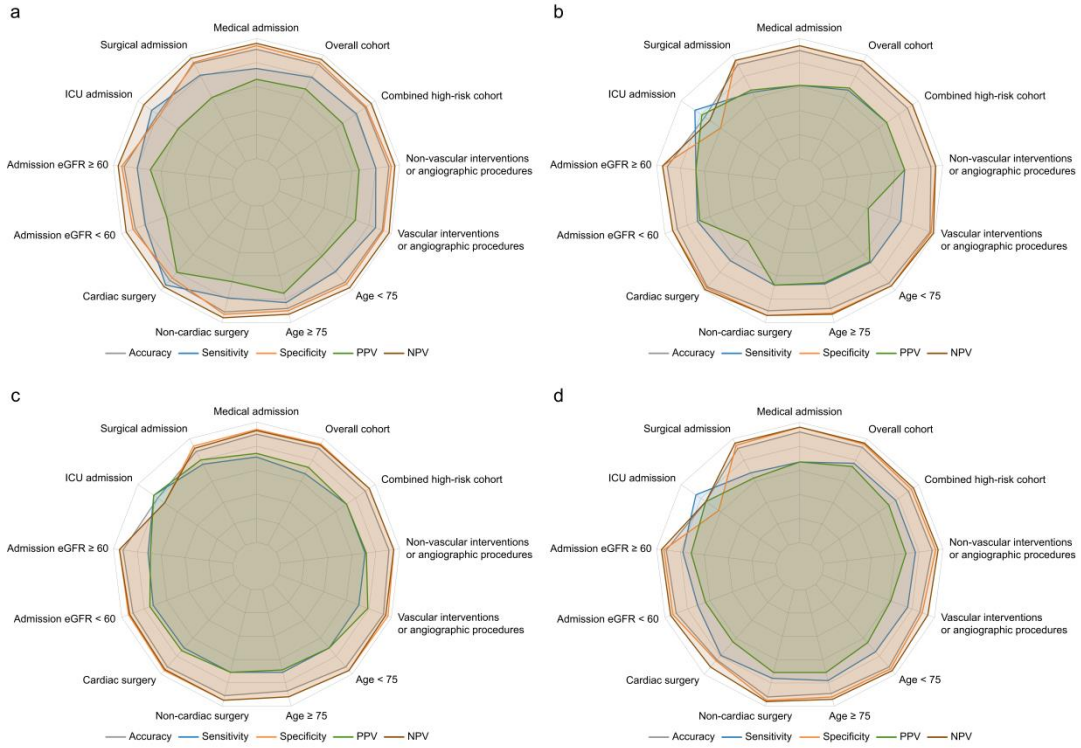
[1] Centers for Disease Control and Prevention (CDC). Identifying Healthcare-Associated Infections (HAIs) for NHSN Surveillance (Chapter 2)[EB/OL]. (January 2026)[2026-04-14]. https://www.cdc.gov/nhsn/PDFs/pscManual/2PSC_IdentifyingHAIs_NHSNcurrent.pdf

7. How will the model perform in an enriched population with high risk? This can be interesting. Like those defined by non-modifiable factors such as advanced age;

Response: We thank the reviewer for this interesting and relevant question. As shown in Figure 3, we conducted stratified analyses across several high-risk subgroups, including patients aged ≥75 years, those undergoing cardiac surgery, vascular interventions or angiographic procedures, ICU admissions, admission eGFR <60 mL/min/1.73m², and a combined high-risk cohort.

Overall, the model showed robust performance in the combined high-risk cohort across both internal and external validation datasets, with no substantial degradation. Specifically, accuracy, specificity, and PPV showed slight decreases of 0.01-0.02, 0.01-0.02, and 0.01-0.04, respectively. In contrast, sensitivity improved marginally by 0.01-0.04, while NPV remained nearly unchanged compared with the overall cohort (see figure below). This consistency was generally maintained across most subgroups, including age ≥75 years, cardiac surgery, vascular interventions or angiographic procedures, and admission eGFR <60 mL/min/1.73m^2. However, the ICU subgroup showed attenuated performance compared with the overall cohort in both internal and external validations: accuracy declined 0.11-0.17, sensitivity improved slightly (increases of 0.07-0.19), whereas specificity declined markedly (reductions of 0.15-0.34). For PPV, we observed a slight decrease in internal validation (down by 0.09), whereas external validation showed small increases of 0.01-0.11. These findings suggest that in the highest-risk ICU setting, the model may favor detection at the expense of more false positives. We have added descriptions and discussed these results and their implications in the Results and Discussion sections, and have provided recommendations for clinical interpretation in different settings (**Page 7-8, line 149-161; Page 14, line 295-299**).

Figure 3. Performance of the AKI-PM for AKI prediction in sub-populations



Performance of AKI-PM for dynamic prediction of AKI within 24 hours with few shot was assessed across sub-populations, stratified by critical clinical factors: age (<75 vs. ≥75 years), admission eGFR (<60 vs. ≥60 ml/min/1.73 m²), admission department (medical, surgical, or ICU), cardiac surgery (yes vs. no), vascular interventions or angiographic procedures (yes vs. no), as well as a high-risk cohort (including patients aged ≥75 years, or those undergoing cardiac surgery, vascular interventions or angiographic procedures, ICU admissions, or those with admission eGFR <60 mL/min/1.73m²). The performance was assessed using accuracy, sensitivity, specificity, PPV, and NPV. Panel a, b, c, and d show the results in PKUFH, Miyun, Taiyuan, and Sichuan, respectively. AKI-PM: acute kidney injury - prediction model; eGFR: estimated glomerular filtration rate; ICU: intensive care unit; PPV: positive predictive value; NPV: negative predictive value; PKUFH: Peking University First Hospital; Miyun: Miyun District Hospital; Taiyuan: Taiyuan Central Hospital; Sichuan: Sichuan Provincial People's Hospital.