

PERIBENCH: A Protocol-Aware Benchmark for Dynamic Perioperative Risk Prediction from Numerical Models to Language Interfaces

Xiaohui Zhao¹ Xiangyu Meng¹ Jinjin Chi^{1,2,*} Yongcheng Jing^{3,*}

Yiming Wang¹ Yibing Zhan^{3,4} Hang Sun⁵ Dacheng Tao²

¹College of Computer Science and Technology, Jilin University

²College of Computing and Data Science, Nanyang Technological University

³School of Computer Science, Wuhan University

⁴Yunnan United Vision Technology Co., Ltd.

⁵College of Computer and Information Technology, China Three Gorges University

xhzha022@mails.jlu.edu.cn chijinjin616@gmail.com yongchengjing@gmail.com

Abstract

Dynamic perioperative risk prediction requires models to anticipate evolving physiological deterioration under strict prediction-time information boundaries. We introduce PERIBENCH, a protocol-aware benchmark built from high-frequency intraoperative signals. PERIBENCH defines 18 clinically motivated tasks spanning future risk classification, persistent deterioration prediction, continuous vital-sign regression, and case-level prognosis, and evaluates 37 systems across numerical predictors, deep sequential models, time-series foundation models, and LLM-oriented language interfaces. Rather than using a single undifferentiated leaderboard, PERIBENCH separates supervised or adapted prediction, direct pretrained transfer, and prompt-based language-interface inference into protocol-compatible pools. Experiments show that model sophistication does not translate monotonically into better perioperative prediction: tree-based references remain strong on many classification endpoints, deep sequential models are competitive on selected physiological tasks, direct foundation-model transfer is heterogeneous, and language interfaces capture useful classification signals but remain less reliable for regression, calibration, alert burden, and abnormal-region prediction. PERIBENCH provides a structured benchmark for studying dynamic perioperative prediction from numerical models to language interfaces under shared task semantics and explicit protocol boundaries. Code is available at <https://github.com/Kdyyh/PERIBENCH>.

1 Introduction

Recent advances in time-series foundation models and large language models have expanded the model space for structured temporal prediction, ranging from numerical predictors and deep sequential models to pretrained time-series models

and language-facing interfaces (Ansari et al., 2024; Jin et al., 2024; Zhang et al., 2024). This diversity creates a benchmark challenge: different systems vary in supervision access, input representation, adaptation strategy, inference protocol, and, for language-facing models, output parsing. For clinical NLP and language-interface evaluation, this setting highlights the role of language interfaces for structured numerical streams through text–numeric serialization, task instruction, and output parsing, rather than only free-text inputs. Fair comparison therefore requires shared task semantics, common metrics, and explicit protocol boundaries (Harutyunyan et al., 2019).

Dynamic perioperative risk prediction is a representative testbed for this benchmark challenge, as prior perioperative resources and studies have shown the value of high-frequency intraoperative signals for physiological prediction (Lee et al., 2022, 2021; Cai et al., 2025). It involves strict prediction-time information boundaries and heterogeneous targets ranging from future risk events and persistent abnormal states to continuous severity values and case-level outcomes. Because these targets differ in temporal semantics, event prevalence, output space, and failure modes, single-endpoint prediction or pure forecasting cannot adequately characterize model capabilities across paradigms, echoing the need for multi-task and protocol-aware clinical time-series evaluation (Harutyunyan et al., 2019; Yèche et al., 2021).

Existing resources only partially address this need. General clinical time-series benchmarks, such as the MIMIC-III benchmark (Harutyunyan et al., 2019) built on MIMIC-III (Johnson et al., 2016), and HiRID-based benchmarks (Yèche et al., 2021), standardize ICU or hospital prediction tasks, but do not target high-frequency perioperative monitoring or cross-paradigm comparison. Perioperative resources such as VitalDB provide valuable intraoperative signals (Lee et al., 2022), and Vi-

*Corresponding authors.

talBench provides a benchmark for intraoperative vital-sign prediction (Cai et al., 2025). However, these resources mainly focus on vital-sign forecasting or selected physiological endpoints, such as intraoperative hypotension prediction (Lee et al., 2021). LLM-based perioperative prediction studies further explore language-facing interfaces (Chung et al., 2024), yet they are often evaluated in isolated prompt settings with limited baselines and inconsistent protocols. These gaps motivate a unified benchmark for numerical, sequential, foundation-model, and language-interface systems under shared dynamic tasks, with protocol-aware evaluation that accounts for differences in input representation and inference interface.

To address this gap, we introduce PERIBENCH, a protocol-aware benchmark for dynamic perioperative risk prediction. Built on high-frequency intraoperative physiological signals, PERIBENCH defines 18 tasks across four prediction regimes: future risk classification, persistent deterioration prediction, continuous vital-sign regression, and case-level prognosis. These regimes cover distinct dynamic prediction semantics and support task-family-level analysis rather than a flat endpoint leaderboard. Beyond task coverage, PERIBENCH is designed for cross-paradigm model evaluation. It evaluates 37 systems spanning numerical predictors, tree-based models, deep sequential networks, time-series foundation models, and LLM-oriented language-interface systems. Since these systems differ in supervision access, adaptation strategy, input representation, and inference interface, PERIBENCH separates them into three protocol-compatible pools: supervised or adapted prediction, direct pretrained transfer, and prompt-based language-interface inference. By explicitly separating systems according to supervision regime, adaptation strategy, and inference protocol, this design preserves shared task definitions and metrics while limiting direct rankings to protocol-compatible systems. Our evaluation shows that no model family is uniformly dominant: performance varies with task semantics, model family, and evaluation protocol.

The contributions of this work are as follows:

- **Task design:** We introduce PERIBENCH, a benchmark with 18 dynamic perioperative risk prediction tasks across four regimes.
- **Cross-paradigm system coverage:** We evaluate 37 systems spanning numerical predic-

tors, tree-based models, deep sequential models, time-series foundation models, and LLM-oriented language-interface systems.

- **Protocol-aware evaluation:** We propose a three-pool design that separates supervised or adapted prediction, direct pretrained transfer, and prompt-based language-interface inference.
- **Reliability analysis:** We provide task-family and reliability-oriented analyses showing that benchmark conclusions depend jointly on task semantics, model family, and evaluation protocol.

2 Data and Task Construction

This section describes how PERIBENCH converts raw intraoperative signals into standardized inputs, task-specific labels, and leakage-controlled evaluation splits.

2.1 Data Source and Standardized Inputs

PERIBENCH is built on VitalDB (Lee et al., 2022) using a unified preprocessing and task-construction pipeline. The original dataset contains 6,388 surgical cases. After excluding 16 cases with corrupted raw data or records that failed the shared preprocessing pipeline, 6,372 cases were retained. For window-level tasks, the basic prediction unit is a 60-second intraoperative physiological window sampled at 1 Hz with a 30-second stride. Each window contains four vital-sign channels: heart rate (HR), mean arterial pressure (MAP), peripheral oxygen saturation (SpO₂), and end-tidal carbon dioxide (ETCO₂). These signals form a tensor of shape $2,181,008 \times 4 \times 60$, paired with a validity mask of the same shape. A unified window manifest of shape $2,181,008 \times 12$ records sample indices, temporal alignment, split assignments, and task labels. Structured case-level variables and laboratory measurements are used only when timestamped before the prediction time. Full task statistics, availability rules, and model-specific input handling are provided in Appendices B and C.

2.2 Task Stratification

The 18 tasks are grouped into four dynamic prediction regimes (Table 1). This stratification is central to PERIBENCH because the regimes capture different prediction semantics rather than merely different labels. Future risk tasks predict threshold

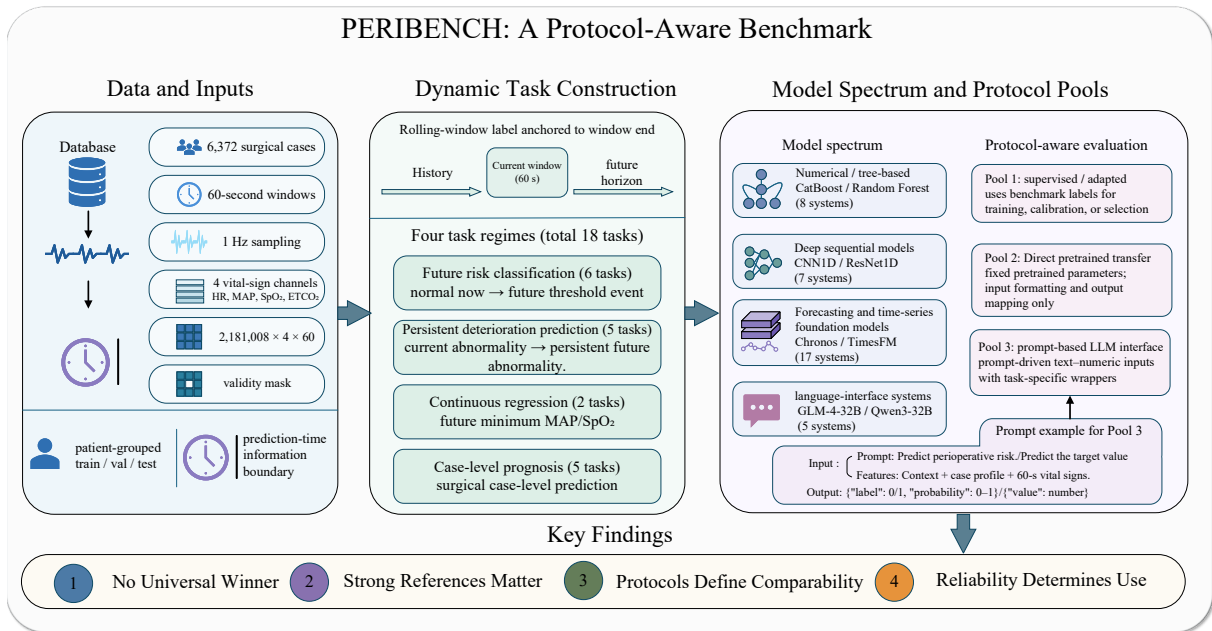


Figure 1: Overview of PERIBENCH. The benchmark is organized along three axes: 18 dynamic prediction tasks, 37 systems spanning numerical predictors, time-series models, foundation models, and language-interface systems, and 3 protocol-stratified evaluation pools. The model spectrum characterizes representation and interface type, while the protocol pool determines ranking comparability.

crossing from a currently stable state; persistent deterioration tasks predict whether an abnormal state will persist; continuous regression tasks estimate future physiological severity; and case-level prognosis tasks summarize episode-level outcomes. Because these regimes differ in sample size, event prevalence, output type, and failure modes, we report results by task family instead of treating all endpoints as a homogeneous label set.

2.3 Rolling-Window Task Construction

Window-level tasks are anchored at the end of each 60-second window, with labels computed over task-specific future horizons. Future risk tasks retain currently stable or normal windows and predict future threshold events, whereas persistent deterioration tasks retain currently abnormal windows and predict whether the abnormal state persists. This separation prevents the two task families from reducing to the same threshold-detection problem. Continuous regression tasks predict the future minimum MAP or SpO₂ over the next 5 minutes. The construction makes the prediction boundary explicit and prevents future observations from entering the input. Detailed endpoint thresholds, state filters, validity requirements, and missingness handling are provided in Appendix D.

2.4 Patient-Grouped Splits and Information Boundaries

All tasks use grouped train, validation, and test partitions at the patient or surgical-case level to prevent leakage across correlated samples from the same trajectory. Model fitting uses the training split, hyperparameter and checkpoint selection use the validation split, and final reporting uses the held-out test split. Task-specific eligible samples inherit the grouped split assignment before label-specific filtering whenever applicable. Window-level labels, case-level covariates, and laboratory measurements are included only when available before the prediction time. This design ensures that benchmark comparisons reflect model behavior under consistent information boundaries rather than access to future observations or patient-level leakage.

3 Model Spectrum

The benchmark evaluates 37 systems across four representation levels (Table 2). The selected systems cover numerical references, sequence architectures, forecasting models, recent time-series foundation models, and language-facing systems. The model spectrum describes representation and interface position, while protocol-compatible ranking is determined separately by the evaluation pool in Section 4. The goal is broad paradigm coverage

Task regime	Count	Definition	Samples	Cases	Event rate	Metrics
Future risk classification	6	stable/normal current window; future threshold crossing	1,266,428–2,122,003	3,521–6,369	2.49–27.86% events	AUROC, AUPRC, Brier
Persistent deterioration prediction	5	abnormal current window; persistent future abnormality	55,736–129,728	1,655–4,233	27.53–66.29% events	AUROC, AUPRC, Brier
Continuous vital-sign regression	2	future minimum MAP or SpO ₂	1,461,123–2,136,422	3,521–6,369	continuous	MAE, RMSE, R^2
Case-level prognosis prediction	5	case-level outcome prediction	3,980–6,372	3,980–6,372	0.86–18.85% events	AUROC, AUPRC, Brier

Table 1: Task stratification in PERIBENCH. Detailed task-level statistics are provided in Appendix B.

rather than an exhaustive enumeration of all available time-series models. Full model references, protocol assignments, and model-specific input protocols are provided in Appendices E and C.

Numerical, feature-based, and tree-based references. The first level contains 8 numerical, feature-based, and tree-based references: Logistic Regression, MLP, MiniRocket, Summary-Feature MLP, Random Forest, XGBoost, LightGBM, and CatBoost. These systems establish strong reference performance using fixed-length features, window summaries, or case-level inputs. They include standard linear and MLP references, random forests (Breiman, 2001), gradient-boosted tree systems (Chen and Guestrin, 2016; Ke et al., 2017; Prokhorenkova et al., 2018), and the MiniRocket time-series transform (Dempster et al., 2020). Their inclusion helps determine whether more complex representation families provide gains over strong established baselines.

Deep sequential models. The second level contains 7 deep sequential models: CNN1D, ResNet1D, InceptionTime, TCN, LSTM, GRU, and Transformer Encoder. These models directly consume multichannel physiological windows and test whether learned temporal representations improve over fixed summary features. They cover recurrent sequence modeling (Hochreiter and Schmidhuber, 1997; Cho et al., 2014), temporal convolution (Bai et al., 2018), residual learning (He et al., 2016), transformer encoders (Vaswani et al., 2017), and time-series-specific convolutional architectures (Ismail Fawaz et al., 2020).

Forecasting and time-series foundation models. The third level contains 17 forecasting and time-series foundation models, covering neural forecasters, transformer- and decomposition-based mod-

els, MLP-style time-series models, and recent pretrained time-series systems. This level tests adaptation, transfer, and direct pretrained inference for perioperative signals. It draws on neural forecasting models (Oreshkin et al., 2020; Challu et al., 2023; Lim et al., 2021), transformer and decomposition-based forecasters (Wu et al., 2021; Zhou et al., 2022; Zeng et al., 2023; Nie et al., 2023; Liu et al., 2024; Wu et al., 2023), MLP-style time-series modeling (Chen et al., 2023), and recent time-series foundation models (Goswami et al., 2024; Ansari et al., 2024; Das et al., 2024; Woo et al., 2024; Rasul et al., 2024; Ekambaram et al., 2024).

LLM-oriented and language-interface systems. The fourth level contains 5 LLM-oriented and language-interface systems: TimeLLM, GPT4MTS, DeepSeek-32B, Qwen3-32B, and GLM-4-32B. TimeLLM is a time-series-to-LLM reprogramming approach (Jin et al., 2024), and GPT4MTS is a task-specific text–numeric modeling implementation (Jia et al., 2024). DeepSeek-32B, Qwen3-32B, and GLM-4-32B are general-purpose LLMs evaluated through language-facing task interfaces and were selected as open-weight 32B checkpoints to enable scale-matched, reproducible local inference under fixed settings rather than exhaustive frontier coverage. Their protocol roles differ: TimeLLM and GPT4MTS are treated as adapted systems, whereas the general-purpose LLMs are evaluated through prompt-based language-interface inference. Together, these systems define the language-facing end of the numerical-to-language-interface spectrum.

4 Protocol-Stratified Evaluation

PERIBENCH uses three protocol-stratified evaluation pools (Table 3). These pools are not intended

Spectrum level	Count	Systems	Protocol assignment
Numerical, feature-based, and tree-based references	8	Logistic Regression, MLP, MiniRocket, Summary-Feature MLP, Random Forest, XGBoost, LightGBM, CatBoost	Pool 1
Deep sequential models	7	CNN1D, ResNet1D, InceptionTime, TCN, LSTM, GRU, Transformer Encoder	Pool 1
Forecasting and time-series foundation models	17	Autoformer, FEDformer, DLinear, PatchTST, iTransformer, TimesNet, Transformer-TS, TSMixer, TFT, N-BEATS, N-HiTS, MOMENT, Chronos, TimesFM, Moirai, Lag-LLaMA, Granite-TTM	12 in Pool 1; 5 in Pool 2
LLM-oriented and language-interface systems	5	TimeLLM, GPT4MTS, DeepSeek-32B, Qwen3-32B, GLM-4-32B	TimeLLM/GPT4MTS: Pool 1; prompt-based LLMs: Pool 3

Table 2: Four-level model spectrum in PERIBENCH. The spectrum describes representation and interface type, while protocol-compatible ranking is defined separately by the evaluation pools. Full model-source references and protocol assignments are provided in Appendix E.

to create a main-versus-secondary hierarchy. Instead, they preserve ranking comparability: the same benchmark can include all 37 systems while avoiding a single mean-rank aggregation across systems with different training, adaptation, or inference assumptions.

Pool 1: Supervised/adapted full-test ranking.

Pool 1 contains 29 systems trained, fitted, or task-adapted using the benchmark training split, selected using the validation split, and evaluated on the held-out full test set. This pool supports within-pool mean-rank aggregation, task-family ranking, and top-system analysis. Model-specific differences in architecture and input formatting remain, but all Pool 1 systems share the same train/validation/test evaluation logic.

Pool 2: Direct pretrained transfer. Pool 2 contains 5 direct pretrained-transfer systems: Chronos, TimesFM, Moirai, Lag-LLaMA, and Granite-TTM. These systems use fixed pretrained parameters without benchmark-label training, parameter updates, calibration, or checkpoint selection. Only input formatting and output-to-task mapping are allowed. Under this protocol, Pool 2 systems are evaluated on temporal-window tasks and are not adapted to the case-level feature set. Pool 2 is reported as direct-transfer performance and compared with Pool 1 references, but is excluded from Pool 1 mean-rank aggregation.

Pool 3: Prompt-based LLM language-interface comparison.

Pool 3 contains three prompt-based LLM language-interface systems: DeepSeek-32B, Qwen3-32B, and GLM-4-32B. They convert benchmark inputs into language-model prediction re-

quests using prompts, text-numeric serialization, and task-specific wrappers. They use neither benchmark-label training nor parameter fine-tuning, but their inference differs from direct pretrained transfer because it depends on prompt design, serialization format, generation settings, and output parsing. Pool 3 results are reported under the same tasks and metrics, but excluded from any shared mean-rank aggregation.

Metrics and reliability-oriented analyses. Classification tasks are evaluated using AUROC, AUPRC, and Brier score, while regression tasks are evaluated using MAE, RMSE, and R^2 . Family-level ranks are computed from task-level ranks using each task’s prespecified primary metric within each protocol-compatible pool, with metric directions normalized. Reliability-oriented analyses further include calibration, F1, sensitivity, specificity, precision, validation-selected operating points, alert burden, threshold-event analysis, and abnormal-region regression. Calibration and operating-point thresholds are selected on the validation split only. Uncertainty is estimated using clustered bootstrap resampling by patient or case group. These reliability-oriented analyses are reported as diagnostics rather than replacements for the primary rank aggregation. Reported sample counts distinguish label-available task cohorts from prespecified interface-eligible subsets defined before inference based on the contiguous-horizon requirements of sequence interfaces; actual evaluated counts may therefore differ across models and protocol pools. Further details on evaluation, rank aggregation, and language-interface prompting and output formats are provided in Appendix F, Ap-

pendix H.1, and Table 22, respectively.

5 Results

This section summarizes the main benchmark results under the protocol-stratified evaluation design. Table 4 reports task-family-level mean ranks, showing the top Pool 1 supervised/adapted systems and all systems in Pools 2 and 3. Ranks are computed only within protocol-compatible pools, and lower ranks indicate better performance. Mean ranks are computed from task-level metric ranks within each task family to summarize heterogeneous endpoints without collapsing incompatible protocols. Classification rank aggregation uses test AUROC, while regression rank aggregation uses test R^2 for benchmark-generated summaries, with metric directions normalized when external error-metric summaries are ranked. The detailed aggregation rule is provided in Appendix H.1. Complete task-level metrics, task-family summaries, and full pool-specific rankings are provided in Appendices G and H.1–H.4.

Pool 1: Supervised/adapted. Pool 1 is the supervised/adapted full-test ranking pool, and its systems should be compared only within this protocol. Table 4 shows that CatBoost has the best Pool 1 overall mean rank (2.808), followed by ResNet1D (4.300), Random Forest (4.833), LightGBM (5.125), and XGBoost (5.517). This ordering is driven mainly by classification endpoints: CatBoost appears in the top five for all 16 classification tasks with mean rank 1.6, while XGBoost and LightGBM appear in the top five for 14/16 tasks and Random Forest for 13/16 tasks. Task-level winners remain endpoint-dependent: CatBoost leads most future-risk endpoints and several case-level outcomes, Random Forest leads persistent bradycardia and hypotension persistence, and ResNet1D leads the low-prevalence future hypoxemia endpoint (AUROC/AUPRC 0.808/0.267).

Continuous regression follows a different pattern. The Pool 1 vital-sign regression rank is led by ResNet1D (2.500), followed by GRU and InceptionTime (4.500 each) and CNN1D (5.500). Metric-level winners are heterogeneous: CNN1D has the lowest MAE for minimum MAP (6.432), CatBoost has the best minimum-MAP RMSE and R^2 (10.397 and 0.568), iTransformer has the lowest MAE for minimum SpO₂ (0.924), and ResNet1D has the best minimum-SpO₂ RMSE and R^2 (2.374 and 0.425). Pool 1 therefore supports the central

result that no universal winner exists even within a single supervised/adapted protocol. Detailed Pool 1 tables are provided in Appendix H.2.

Pool 2: Direct pretrained transfer. Pool 2 contains direct pretrained-transfer systems that do not use benchmark-label training and do not update parameters. These systems are therefore compared within Pool 2 and should not be merged into the Pool 1 leaderboard. Table 4 shows that Chronos has the best Pool 2 overall mean rank (1.700), followed by TimesFM (2.356), Moirai (2.544), Granite-TTM (3.600), and Lag-LLaMA (4.800). No Pool 2 system is reported for case-level prognosis under our direct-transfer protocol, because these pretrained forecasting interfaces are evaluated only on temporal-window tasks and are not adapted to the case-level feature set.

Transfer behavior is heterogeneous across task families. Chronos has the best Pool 2 future-risk rank (1.500) and vital-sign regression rank (1.000), while Moirai has the best persistent-deterioration rank (1.800). Chronos leads all three regression metrics on both continuous tasks, with minimum MAP MAE/RMSE/ R^2 of 8.243/13.749/0.238 and minimum SpO₂ values of 0.794/2.992/−0.048. Although Chronos ranks first within Pool 2, its negative R^2 on minimum SpO₂ indicates weak absolute regression quality for that endpoint. The much weaker Lag-LLaMA regression values, including minimum-MAP MAE 72.584 and R^2 = −664.022, further show that direct transfer can provide useful signal for selected tasks but is not a uniformly reliable substitute for supervised/adapted ranking. Detailed Pool 2 results are provided in Appendix H.3.

Pool 3: Prompt-based language interface. Pool 3 is the prompt-based LLM language-interface comparison. It relies on prompt-driven text–numeric inference, input serialization, and output parsing, so its results must not be merged with Pool 1 or Pool 2 into a shared leaderboard. Table 4 reports two Pool 3 input settings. The temporal-window setting serializes the 60-second physiological window together with the task description and available case profile, whereas the no-temporal-window setting keeps the same task description, case-profile fields, output schema, and parsing rules but removes the raw 60-second vital-sign sequence. The case-profile fields follow the same prediction-time availability rules as the numerical systems. With temporal-window input, Qwen3-

Protocol pool	#	Protocol boundary	Ranking role
Pool 1: Supervised/adapted	29	Benchmark labels are used for training, fitting, fine-tuning, task adaptation, calibration, hyperparameter selection, or checkpoint selection.	Supports direct within-pool ranking, mean-rank aggregation, task-family ranking, and top-model analysis.
Pool 2: Direct pretrained transfer	5	Fixed pretrained parameters are used without benchmark-label training or parameter updates; only input formatting and output mapping are allowed.	Compared against Pool 1 references as direct transfer results; excluded from Pool 1 mean ranks.
Pool 3: Prompt-based LLM interface	3	General-purpose LLMs are evaluated through prompt-driven text–numeric interfaces and task-specific wrappers.	Used only for protocol-aware language-interface comparison; excluded from shared mean-rank aggregation.

Table 3: Protocol-stratified evaluation pools. All pools use the same benchmark tasks and test labels, but ranking comparability is determined by how models access benchmark supervision and how inference is performed; direct ranking is defined only within protocol-compatible pools.

System	System type	Future risk	Persistent det.	Vital-sign reg.	Case-level prog.	Overall rank
Pool 1: supervised / adapted full-test ranking						
CatBoost	Tree-based	1.333	<u>2.000</u>	6.500	1.400	2.808
ResNet1D	Deep sequential	4.000	6.400	2.500	–	<u>4.300</u>
Random Forest	Tree-based	5.333	1.400	10.000	<u>2.600</u>	4.833
LightGBM	Tree-based	4.000	3.600	9.500	3.400	5.125
XGBoost	Tree-based	<u>2.667</u>	3.800	13.000	<u>2.600</u>	5.517
GRU	Deep sequential	8.333	6.200	<u>4.500</u>	–	6.344
InceptionTime	Deep sequential	5.667	9.200	<u>4.500</u>	–	6.456
CNN1D	Deep sequential	8.333	9.000	5.500	–	7.611
LSTM	Deep sequential	8.833	11.600	6.000	–	8.811
Transformer Encoder	Deep sequential	11.667	9.600	8.000	–	9.756
<i>Remaining 19 Pool 1 systems are reported in Appendix G</i>		...	...	...	...	...
Pool 2: direct pretrained-transfer full-test comparison						
Chronos	TS foundation / direct	1.500	2.600	1.000	–	1.700
TimesFM	TS foundation / direct	2.667	<u>2.400</u>	<u>2.000</u>	–	<u>2.356</u>
Moirai	TS foundation / direct	<u>1.833</u>	1.800	4.000	–	2.544
Granite-TTM	TS foundation / direct	4.000	3.800	3.000	–	3.600
Lag-LLaMA	TS foundation / direct	5.000	4.400	5.000	–	4.800
Pool 3: prompt-based LLM language-interface comparison						
<i>With temporal-window input</i>						
DeepSeek-32B	LLM interface + TS	3.000	<u>1.800</u>	1.500	2.800	2.275
Qwen3-32B	LLM interface + TS	<u>1.667</u>	1.600	<u>2.000</u>	<u>2.000</u>	1.817
GLM-4-32B	LLM interface + TS	1.333	2.600	2.500	1.200	<u>1.908</u>
<i>no-temporal-window</i>						
DeepSeek-32B	LLM interface	2.500	2.400	3.000	3.000	2.725
Qwen3-32B	LLM interface	<u>2.000</u>	1.400	<u>2.000</u>	<u>2.000</u>	<u>1.850</u>
GLM-4-32B	LLM interface	1.500	<u>2.200</u>	1.000	1.000	1.425

Table 4: Protocol-aware family-level mean ranks in PERIBENCH. Lower is better. Ranks are computed within protocol-compatible pools and are not comparable across pools. Pool 1 shows the top ten supervised/adapted systems, while Pool 2 and Pool 3 include all systems. Full Pool 1 results are provided in Appendix G. Bold and underlined values mark the best and second-best ranks within each protocol pool; for Pool 3, they are computed separately within each input setting. Dashes indicate that the system was not evaluated for that task family under the corresponding protocol.

32B achieves the best overall mean rank (1.817), followed by GLM-4-32B (1.908) and DeepSeek-32B (2.275). In the no-temporal-window setting, GLM-4-32B achieves the best overall mean

rank (1.425), followed by Qwen3-32B (1.850) and DeepSeek-32B (2.725). Because Pool 3 contains only three language-interface systems, its ranks should be interpreted as diagnostic comparisons rather than broad model-family rankings.

The language-interface pool retains meaningful classification signal, but the leading system changes by task family and input setting. With temporal-window input, GLM-4-32B has the highest mean future-risk AUROC/AUPRC (0.693/0.246) and the highest case-level mean AUROC (0.797), while Qwen3-32B has the highest persistent-deterioration mean AUROC/AUPRC (0.607/0.522). Continuous regression is more sensitive to serialization: Qwen3-32B leads minimum-MAP regression with temporal input (MAE/RMSE/ R^2 7.166/13.140/0.277), DeepSeek-32B leads minimum-SpO₂ with temporal input (0.773/2.637/0.394), and DeepSeek-32B degrades sharply in the no-temporal-window setting (R^2 = -0.354 for MAP and -0.744 for SpO₂). These results highlight that language-interface evaluation depends not only on the LLM endpoint, but also on how structured time series are serialized. Detailed Pool 3 results, prompt templates, and output-parsing rules are provided in Appendix H.4.

Targeted robustness analyses. In a targeted evaluation on an independent UQ cohort, we assessed the three Pool 3 language-interface systems on future hypoxemia, future tachycardia, and minimum-SpO₂ regression (Table 5, Panel A). Qwen3-32B achieved the highest hypoxemia AUROC (0.617), tachycardia AUROC/AUPRC (0.660/0.206), and minimum-SpO₂ RMSE (2.701), whereas GLM-4-32B achieved the highest hypoxemia AUPRC (0.042) and the lowest minimum-SpO₂ MAE (1.154). No model was best on every metric, and performance remained heterogeneous across endpoints and models. In complementary input ablations, the best-performing configuration differed across models for future-hypotension classification, whereas removing the raw sequence increased minimum-SpO₂ MAE for all three systems (Table 5, Panel B).

Reliability beyond mean ranks. Aggregate metrics and mean ranks are insufficient for reliability-sensitive interpretation. The following analyses do not replace the primary ranks, but expose reliability failures that aggregate scores may hide. Calibration results show that strong discrimination can coexist with different probability-error profiles. CatBoost

Panel A. External evaluation					
Model	Hypoxemia	Tachycardia	Minimum SpO ₂		
DeepSeek-32B	0.545/ <u>0.038</u>	0.627/0.158	1.157/2.830		
Qwen3-32B	0.617 /0.030	0.660 / 0.206	1.162/ 2.701		
GLM-4-32B	<u>0.582</u> / 0.042	0.561/0.154	1.154 /2.943		
Panel B. Input ablation					
Model	Full	No raw	Summ.	Raw only	Profile only
Future hypotension					
DeepSeek-32B	0.521	<u>0.558</u>	0.574	0.508	0.488
Qwen3-32B	0.606	0.632	<u>0.613</u>	0.567	0.500
GLM-4-32B	0.681	<u>0.658</u>	0.632	0.585	0.573
Minimum SpO ₂					
DeepSeek-32B	0.777	<u>3.506</u>	—	—	—
Qwen3-32B	0.885	<u>1.756</u>	—	—	—
GLM-4-32B	0.772	<u>1.491</u>	—	—	—

Table 5: Targeted robustness analyses. Panel A reports AUROC/AUPRC and MAE/RMSE; Panel B reports AUROC and MAE. Best and second-best results are bold and underlined.

has strong discrimination and low calibration error for future bradycardia (AUROC 0.915, AUPRC 0.485, Brier 0.030, ECE 0.005) and ICU admission (AUROC 0.892, AUPRC 0.717, Brier 0.090, ECE 0.023), whereas Random Forest on persistent hypotension at 5 minutes has AUROC 0.754 and AUPRC 0.689 but a larger ECE of 0.057. Operating thresholds selected on the validation split and applied to the held-out test set further show that high discrimination does not imply low alert burden: at a 95% specificity target, ResNet1D on future hypoxemia reaches sensitivity 0.484 and PPV 0.214 but produces 0.884 false alerts per eligible monitoring hour and 4.958 alerts per case.

Regression analyses require the same reliability-oriented interpretation. Threshold-event analysis for MAP < 65 mmHg identifies LightGBM as the best AUROC model (0.878) and ResNet1D as the best AUPRC model (0.816). For SpO₂ < 90%, CatBoost has the best threshold-event AUROC (0.840) and ResNet1D has the best AUPRC (0.356). Abnormal-region errors differ from global errors: CatBoost has overall minimum-MAP MAE 6.584 but true-below-65 MAE 10.403 and miss rate 0.270; on the shared eligible subset for minimum SpO₂, CatBoost has overall MAE 1.296 but true-below-90 MAE 10.579 and miss rate 0.684. Extended reliability tables are provided in Appendix H.5.

Paired ranking uncertainty. We further examined the stability of the AUROC differences between the two leading Pool 1 systems for four representative classification tasks. Paired case-group-clustered bootstrap comparisons with 1,000

Task	Models	AUROC	Δ (95% CI)	Holm p
Future bradycardia, 5 min	CatBoost/XGBoost	0.915/0.912	0.0028 (0.0009, 0.0046)	0.032
Future hypoxemia, 5 min	ResNet1D/CatBoos	0.808/0.801	0.0066 (−0.0028, 0.0163)	0.364
Persistent hypotension, 5 min	Random Forest/CatBoost	0.754/0.751	0.0030 (−0.0099, 0.0162)	0.695
ICU admission	CatBoost/LightGBM	0.892/0.882	0.0100 (0.0009, 0.0190)	0.096

Table 6: Paired case-group-clustered bootstrap comparisons for four representative Pool 1 tasks. Differences are calculated as the first-listed model minus the second-listed model.

resamples and Holm correction showed that only the future-bradycardia difference remained significant after correction (Δ AUROC = 0.0028, 95% CI [0.0009, 0.0046], adjusted p = 0.032; Table 6). The differences for future hypoxemia, persistent hypotension, and ICU admission were not significant after correction, indicating that close point estimates among leading systems do not necessarily represent stable performance differences across case groups.

6 Discussion

No universal winner. PERIBENCH shows that dynamic perioperative prediction cannot be summarized by a single best architecture. The four task regimes emphasize different capabilities: early event anticipation, persistence modeling, numerical severity estimation, and case-level aggregation. This explains why tree ensembles remain strong on many classification endpoints, while deep sequential systems and direct pretrained-transfer models are useful only for selected physiological or continuous prediction settings.

Strong references matter. Numerical, feature-based, and tree-based references are essential for interpreting newer systems. CatBoost, Random Forest, XGBoost, and LightGBM set strong classification standards, preventing deep models, foundation models, or language-interface systems from appearing stronger merely because of weak baselines. Their inclusion helps separate representation gains from protocol effects and task-specific feature stability.

Protocols define comparability. Model spectrum and ranking comparability are distinct. TimeLLM and GPT4MTS are LLM-oriented systems but belong to Pool 1 because they

use benchmark-label training or task adaptation, whereas DeepSeek-32B, Qwen3-32B, and GLM-4-32B belong to Pool 3 as prompt-based language-interface systems. Protocol-stratified evaluation keeps shared tasks and metrics while avoiding the merger of incompatible supervision, adaptation, and inference assumptions into one leaderboard. Pool 3 further shows that language-interface performance varies with model choice, task semantics, and input serialization, with temporal information providing task-dependent benefits.

Reliability determines use. Aggregate metrics and mean ranks are insufficient for reliability-sensitive evaluation. Calibration, validation-selected operating points, clustered uncertainty, alert burden, threshold-event analysis, and abnormal-region errors expose failures that aggregate scores may hide. These checks are especially important for low-MAP and low-SpO₂ regions and for language-interface systems, where plausible outputs may still lack numerical precision or stable threshold behavior.

Implications for clinical NLP. Clinical NLP researchers can use PERIBENCH to study how text-numeric representations and output interfaces affect prediction and reliability under shared task semantics; it should not support cross-pool rankings or claims of clinical readiness or cross-institutional validity.

7 Conclusion

Across 18 tasks and 37 systems organized into 3 protocol-stratified evaluation pools, no architecture is universally optimal. PERIBENCH shows that task semantics, model representation, and protocol comparability jointly shape benchmark conclusions. Strong numerical, feature-based, and tree-based references remain difficult to beat on many classification tasks; deep sequential systems are valuable for selected physiological endpoints; time-series foundation and direct pretrained-transfer systems show heterogeneous behavior rather than universal gains; and language-interface systems require protocol-aware interpretation. PERIBENCH therefore provides a structured framework for evaluating dynamic perioperative prediction without collapsing systems with incompatible supervision, adaptation, and inference assumptions into a single undifferentiated leaderboard.

Limitations

This study has several limitations. First, PERIBENCH is constructed from VitalDB, and the targeted external evaluation covers only three signal-compatible endpoints in one independent UQ cohort; broader validation across institutions, surgical populations, monitoring devices, and prospective workflows remains necessary. Second, the tasks use standardized physiological thresholds and future observation windows; benchmark performance therefore reflects prediction of reproducible targets rather than evidence of improved clinical outcomes. Third, protocol stratification improves fairness but prevents a single global ranking across all systems. Fourth, the language-interface evaluation includes only three general-purpose LLMs and depends on prompt templates, text–numeric serialization, generation settings, and deterministic output parsing; broader model coverage and prompt-sensitivity analyses remain future work. Offline calibration, alert-burden, and abnormal-region analyses also cannot replace prospective evaluation with clinician interaction and patient-outcome assessment.

Ethics and Data Privacy

This work uses VitalDB, a publicly available and anonymized perioperative research database. The VitalDB documentation reports that the acquisition and release of the data were approved by the Institutional Review Board of Seoul National University Hospital (IRB No. H-1408-101-605), and that the study was registered at ClinicalTrials.gov (NCT02914444). Our study is limited to secondary analysis of the released research dataset. We do not recruit patients, contact individuals, perform clinical interventions, or collect additional identifiable information. To respect the source-data access policy, PERIBENCH will release code, task registries, split manifests, model configurations, prediction files, evaluation scripts, and language-interface artifacts rather than redistributing restricted raw records. Users with authorized VitalDB access will be able to reconstruct the benchmark locally. PERIBENCH is designed for retrospective benchmark evaluation and should not be interpreted as a clinically deployed risk prediction system. Any prospective clinical use would require separate institutional, ethical, and regulatory review.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments, which helped improve this paper. This work was supported by the Key Research and Development Program of the Jilin Provincial Department of Science and Technology under Grant No. 20260202034NC and the Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, China.

References

- Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. 2024. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*.
- Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.
- Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.
- Xiuding Cai, Xueyao Wang, Sen Wang, Yaoyao Zhu, Jiao Chen, and Yu Yao. 2025. [Vitalbench: A rigorous multi-center benchmark for long-term vital sign prediction in intraoperative care](#). *Preprint*, arXiv:2511.13757.
- Cristian Challu, Kin G. Olivares, Boris N. Oreshkin, Federico Garza, Max Mergenthaler-Canseco, and Artur Dubrawski. 2023. N-HiTS: Neural hierarchical interpolation for time series forecasting. In *AAAI conference on artificial intelligence*, volume 37, pages 6989–6997.
- Si-An Chen, Chun-Liang Li, Nate Yoder, Sercan O. Arik, and Tomas Pfister. 2023. TSMixer: An all-MLP architecture for time series forecasting. *arXiv preprint arXiv:2303.06053*.
- Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Empirical Methods in Natural Language Processing*, pages 1724–1734.

- Philip Chung, Christine T. Fong, Andrew M. Walters, Nima Aghaeepour, Meliha Yetisgen, and Vikas N. O'Reilly-Shah. 2024. Large language model capabilities in perioperative risk prediction and prognostication. *JAMA Surgery*, 159(8):928–937.
- David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20(2):215–232.
- Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. 2024. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 10148–10167. PMLR.
- Angus Dempster, Daniel F. Schmidt, and Geoffrey I. Webb. 2020. MINIROCKET: A very fast (almost) deterministic transform for time series classification. *arXiv preprint arXiv:2012.08791*.
- Vijay Ekambaram, Arindam Jati, Pankaj Dayama, Sumanta Mukherjee, Nam H. Nguyen, Wesley M. Gifford, Chandra Reddy, and Jayant Kalagnanam. 2024. Tiny time mixers (TTMs): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. In *Neural Information Processing Systems*.
- Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. 2024. MO-MENT: A family of open time-series foundation models. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 16115–16152. PMLR.
- Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, pages 770–778.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.
- Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. 2020. InceptionTime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34:1936–1962.
- Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. 2024. Gpt4mts: Prompt-based large language model for multimodal time-series forecasting. In *AAAI conference on artificial intelligence*, volume 38, pages 23343–23351.
- Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2024. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*.
- Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-Wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035.
- Nari Johnson, Sonali Parbhoo, Andrew S Ross, and Finale Doshi-Velez. 2022. Learning predictive and interpretable timeseries summaries from icu data. In *AMIA Annual Symposium Proceedings*, volume 2021, page 581.
- Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. In *Neural Information Processing Systems*, volume 30.
- Hyung-Chul Lee, Yoonsang Park, Soo Bin Yoon, Seong Mi Yang, Dongnyeok Park, and Chul-Woo Jung. 2022. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9(1):279.
- Sangmin Lee, Hyung-Chul Lee, Youn Joung Chu, Su Jeong Song, Jae Hwan Lee, Hee Yeon Kim, Jong-Hwan Ahn, Sohee Yang, Seung Bin Koh, Sang-Hwan Lee, and Byung-Moon Choi. 2021. Deep learning models for the prediction of intraoperative hypotension. *British Journal of Anaesthesia*, 126(4):808–817.
- Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. 2021. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764.
- Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. 2024. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*.
- Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*.
- Boris N. Oreshkin, Dmitri Carпов, Nicolas Chapados, and Yoshua Bengio. 2020. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*.
- Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey

- Gulin. 2018. CatBoost: Unbiased boosting with categorical features. In *Neural Information Processing Systems*, volume 31.
- Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Bilos, Hena Ghonia, Nadhir Hassen, Anderson Schneider, Sahil Garg, Alexandre Drouin, Nicolas Chapados, Yuriy Nevmyvaka, and Irina Rish. 2024. Lag-llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv:2310.08278*.
- David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature*, 323(6088):533–536.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*, volume 30.
- Zhiguang Wang, Weizhong Yan, and Tim Oates. 2017. Time series classification from scratch with deep neural networks: A strong baseline. In *International joint conference on neural networks*, pages 1578–1585.
- Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. 2024. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 53140–53164. PMLR.
- Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. 2023. TimesNet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*.
- Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Neural Information Processing Systems*, volume 34, pages 22419–22430.
- Hugo Yèche, Rita Kuznetsova, Marc Zimmermann, Matthias Hüser, Xinrui Lyu, Martin Faltys, and Gunnar Rätsch. 2021. Hirid-icu-benchmark: A comprehensive machine learning benchmark on high-resolution icu data. In *Neural Information Processing Systems Track on Datasets and Benchmarks*.
- Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. 2023. Are transformers effective for time series forecasting? In *AAAI conference on artificial intelligence*, volume 37, pages 11121–11128.
- Xiyuan Zhang, Ranak Roy Chowdhury, Rajesh K Gupta, and Jingbo Shang. 2024. Large language models for time series: a survey. In *International Joint Conference on Artificial Intelligence*, pages 8335–8343.
- Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. 2022. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27268–27286. PMLR.

A Appendix Overview

This appendix provides supplementary details for task statistics, model-specific input protocols, window-level task construction, model references, detailed evaluation protocols, protocol-stratified ranking, extended pool-specific results, clinical reliability analyses, LLM prompt construction, and machine-readable result files.

- Appendix B: task-level cohort statistics in Table 7.
- Appendix C: model-specific input construction and evaluation handling in Table 9.
- Appendix D: window-level task construction, including prediction boundaries, endpoint thresholds, state filtering, label construction, validity filtering, missingness handling, and final task counts in Tables 10 and 11.
- Appendix E: model zoo, protocol pools, and model-source references in Table 12.
- Appendix F: detailed evaluation protocol, including classification metrics, regression metrics, clustered uncertainty, calibration, operating points, alert burden, and abnormal-region regression.
- Appendix G: complete protocol-stratified family-level ranking in Table 13.
- Appendix H.1: mean-rank aggregation procedure used to summarize heterogeneous endpoint performance across protocol-compatible pools.
- Appendix H.2: detailed supervised/adapted Pool 1 results, including family-level ranks, task-level classification winners, and regression leaders in Tables 14–16.
- Appendix H.3: detailed direct pretrained-transfer Pool 2 results, including family-level ranks, classification winners, and regression metrics in Tables 17–19.
- Appendix H.4: detailed Pool 3 prompt-based LLM language-interface results, including temporal-window and no-temporal-window settings, regression summaries, and prompt/output parsing rules in Tables 20–22.
- Appendix H.5: extended clinical reliability analyses, including calibration, operating-point behavior, alert burden, threshold-event analyses, abnormal-region regression, and clustered bootstrap confidence intervals in Tables 23–28.
- Appendix I: planned code release and reproducibility package, including data reconstruction scripts, task registry, split manifests, model configurations, prediction files, evaluation code, reliability-analysis utilities, language-interface components, and leaderboard rules in Table 29.

B Task Details and Statistics

Table 7 reports the full task-level cohort statistics for the 18 benchmark tasks. Each row lists task type, task group, sample or window count, cases, train/validation/test counts, and event rate or target summary.

C Model-Specific Input Protocols

All window-level tasks use the same 60-second, four-channel physiological window and validity mask. Different models consume this shared source as tabular summaries, sequence tensors, forecasting contexts, text–numeric inputs, or prompts. Case-level endpoints use only variables and summaries available before the prediction boundary. This appendix summarizes model-specific input construction and evaluation handling under the protocol-stratified design in Section 4. Table 9 provides the detailed breakdown of model-specific input construction and evaluation handling.

D Window-level Task Construction

D.1 Prediction Boundary and Endpoint Definitions

All window-level tasks are anchored at the end of each 60-second input window. For a window ending at time t , only observations within the input window are used as model input, while labels are computed from future observations in $(t, t + H]$, where H denotes the task-specific prediction horizon. This construction makes the prediction-time boundary explicit and prevents future observations from entering the input.

Let $A_e(x)$ denote whether endpoint e is abnormal under its clinical threshold. Future-risk tasks retain windows that are currently stable/normal and

Task	Type	Group	Samples	Cases	Train/Val/Test	Event rate
Future bradycardia, 5 min	classification	future risk binary	2,029,350	6,369	1,435,711/201,997/391,642	4.40%
Future hypotension, 5 min	classification	future risk binary	1,287,334	3,521	910,327/128,381/248,626	17.60%
Future hypotension, 10 min	classification	future risk binary	1,266,428	3,521	895,638/126,259/244,531	27.86%
Future hypoxemia, 5 min	classification	future risk binary	2,122,003	6,368	1,501,246/211,221/409,536	2.49%
Future low ET _{CO} ₂ , 5 min	classification	future risk binary	1,932,620	6,141	1,366,823/192,570/373,227	10.88%
Future tachycardia, 5 min	classification	future risk binary	1,968,380	6,330	1,392,266/196,153/379,961	13.05%
Persistent bradycardia, 5 min	classification	persistent binary	55,736	1,655	39,388/5,575/10,773	51.98%
Persistent hypotension, 5 min	classification	persistent binary	129,728	3,009	91,750/12,913/25,065	39.67%
Persistent hypotension, 10 min	classification	persistent binary	129,600	3,016	91,675/12,909/25,016	27.53%
Persistent low ET _{CO} ₂ , 5 min	classification	persistent binary	81,234	4,233	57,463/8,090/15,681	42.58%
Persistent tachycardia, 5 min	classification	persistent binary	105,072	3,106	74,318/10,459/20,295	66.29%
Minimum MAP, next 5 min	regression	regression	1,461,123	3,521	1,032,384/146,340/282,399	mean 72.341, SD 15.603
Minimum SpO ₂ , next 5 min	regression	regression	2,136,422	6,369	1,510,672/213,081/412,669	mean 98.565, SD 3.285
In-hospital death	classification	case	6,372	6,372	4,508/634/1,230	0.86%
ICU admission	classification	case	6,372	6,372	4,507/635/1,230	18.85%
Estimated blood loss ≥500 mL	classification	case	3,980	3,980	2,816/396/768	16.38%
Estimated blood loss ≥1000 mL	classification	case	3,980	3,980	2,816/396/768	6.13%
Any transfusion	classification	case	6,372	6,372	4,508/635/1,229	5.65%

Table 7: Full task-level statistics.

predict whether a future threshold event occurs. Persistent-deterioration tasks retain windows that are already stably abnormal and predict whether the abnormal state persists into the future horizon. Continuous regression tasks predict the minimum valid future MAP or SpO₂ value over the next 5 minutes. The endpoint definitions and abnormal thresholds are listed in Table 10.

D.2 State Filtering and Label Construction

For current-state filtering, the last 10 seconds of the input window are used as the current tail. A tail is eligible only if at least 80% of its samples are valid after physiological quality control. For endpoint e , a window is considered stable/normal if the abnormal fraction in the current tail is at most 0.2, and stably abnormal if the abnormal fraction is at least

0.8. Future-risk tasks retain the stable/normal windows, whereas persistent-deterioration tasks retain the stably abnormal windows. Thus, the two task families are constructed from disjoint current-state regimes rather than reducing to the same future threshold-detection problem.

For a future-risk task with horizon H , a retained window receives a positive label if any valid future sample in $(t, t + H]$ crosses the endpoint-specific abnormal threshold:

$$y_{\text{risk}} = \mathbb{I} \left(\frac{1}{|\mathcal{F}|} \sum_{\tau \in \mathcal{F}} A_e(x_\tau) > 0 \right),$$

where $\mathcal{F}$ denotes the set of valid future samples. The future segment must have at least 80% valid coverage. The benchmark defines 5-minute future-risk tasks for hypoxemia, bradycardia, tachycardia,

Task family	Representative input	Prediction label or target	Metrics	Eligible pools
Future risk classification	Current 60-s, 1-Hz physiological window containing HR, MAP, SpO ₂ , and ETCO ₂ , together with its validity mask; represented according to the model protocol	Endpoint-specific threshold crossing within the future horizon, conditional on a currently stable or normal state	AUROC, AUPRC, Brier	P1 P2 P3
Persistent deterioration prediction	Current 60-s physiological window restricted to samples with a stably abnormal current tail; represented according to the model protocol	Sustained abnormality over the future horizon with the endpoint remaining abnormal at the horizon boundary	AUROC, AUPRC, Brier	P1 P2 P3
Continuous vital-sign regression	Current 60-s, four-channel physiological window and validity mask, represented as numerical features, a temporal sequence, a pretrained-model input, or a text-numeric prompt	Minimum valid MAP or SpO ₂ value observed over the next 5 minutes	MAE, RMSE, R^2	P1 P2 P3
Case-level prognosis prediction	Permitted case-profile variables, timestamped laboratory measurements, and pooled physiological summaries available before the task-specific prediction boundary	In-hospital death, ICU admission, estimated blood loss thresholds, or any transfusion	AUROC, AUPRC, Brier	P1 P3

Table 8: Mapping of PERIBENCH task families to representative inputs, prediction targets, evaluation metrics, and eligible protocol pools. P1 denotes supervised or adapted evaluation, P2 denotes direct pretrained transfer, and P3 denotes prompt-based language-interface evaluation. Pool eligibility reflects task and input compatibility; rankings remain restricted to protocol-compatible pools.

low ETCO₂, and hypotension, and additionally includes a 10-minute future hypotension task.

For persistent-deterioration tasks, the retained window is already stably abnormal at the prediction boundary. A positive label requires both sustained abnormality over the future horizon and abnormality at the horizon endpoint:

$$y_{\text{persist}} = \mathbb{I} \left(\frac{1}{|\mathcal{F}|} \sum_{\tau \in \mathcal{F}} A_e(x_\tau) \geq 0.8 \right) \cdot \mathbb{I}(A_e(x_{t+H}) = 1).$$

This definition excludes windows in which the abnormal state rapidly normalizes before the end of the horizon. Persistent-deterioration tasks are defined for hypotension at 5 and 10 minutes, and for bradycardia, tachycardia, and low ETCO₂ at 5 minutes.

For continuous regression tasks, no binary abnormal label is used. The target is the minimum valid future value over the next 5 minutes:

$$y_{\text{MAP}} = \min_{\tau \in (t, t+5\text{min}]} \text{MAP}_\tau,$$

$$y_{\text{SpO}_2} = \min_{\tau \in (t, t+5\text{min}]} \text{SpO}_{2,\tau}.$$

Regression windows are retained only when the future horizon has at least 80% valid coverage after physiological quality control.

D.3 Validity Filtering, Missingness Handling, and Task Counts

Physiological quality-control ranges are applied before computing coverage, current-state filters, and labels: HR must be within 20–250 beats/min, MAP within 20–200 mmHg, SpO₂ within 40–100%, and ETCO₂ within 5–120 mmHg. The task registry records a maximum hold interval of 10 seconds. Windows that do not meet the current-tail or future-horizon coverage requirements after this validity procedure are excluded.

Missingness and invalid measurements are handled with binary validity masks. Each numerical input window is paired with a validity mask. Raw missing samples and samples failing physiological quality control are marked as invalid, excluded from coverage calculation, state filtering, and label construction, and set to zero in the numerical tensor. The mask distinguishes these zero-filled invalid entries from true observed physiological values. Only valid future samples are used to compute future abnormal fractions, terminal abnormality, and regression targets.

All canonical tasks are split by subject group, with no subject overlap across training, validation, and test splits. Table 11 reports the final window-level task counts.

Protocol pool	Model family	Input construction	Evaluation handling
Pool 1	Supervised tabular and tree references	Engineered window summaries, including value, trend, variability, validity, missingness, and abnormal-fraction features; case-level tasks use permitted case-level variables and pooled summaries.	Trained on the training split, selected with the validation split, evaluated on the full test split, and included in Pool 1 ranking.
Pool 1	Supervised deep sequence models	Four-channel 1 Hz, 60-second tensors with validity-mask handling through model-specific masking, normalization, or imputation.	Optimized with benchmark labels; validation-selected checkpoints are evaluated on the full test split and included in Pool 1.
Pool 1	Supervised forecasting and time-series adaptation models	Physiological windows are reformatted into model-specific sequence, forecasting, or representation-learning inputs; outputs are mapped to classification or regression targets.	Adapted or trained with benchmark labels or future targets, and treated as supervised/adapted Pool 1 models.
Pool 1	Adapted language-conditioned local models	TimeLLM and GPT4MTS use local language-conditioned or text-numeric wrappers derived from benchmark windows, case text, numeric summaries, or raw sequences.	Trained or adapted with benchmark labels and included in Pool 1 when reported as finalized benchmark results; exploratory variants are kept separate unless explicitly merged.
Pool 2	Direct pretrained-transfer time-series models	Chronos, TimesFM, Moirai, Lag-LLaMA, and Granite-TTM use pretrained forecasting interfaces with input formatting and task-specific output reduction only.	No benchmark-label training or parameter update; evaluated on the same full test tasks but reported separately from Pool 1 ranking.
Pool 3	Prompt-based language-interface models	DeepSeek-32B, Qwen3-32B, GLM-4-32B, and similar prompt-based models serialize benchmark inputs into prompts and parse generated outputs into task predictions.	No benchmark-label training or fine-tuning; reported as a protocol-distinct language-interface pool and excluded from unified cross-pool ranking.

Table 9: Model-specific input construction and evaluation handling under the protocol-stratified benchmark design.

Endpoint	Signal	Abnormal threshold
Hypotension	MAP	MAP < 65 mmHg
Hypoxemia	SpO ₂	SpO ₂ < 92%
Bradycardia	Heart rate	HR < 50 beats/min
Tachycardia	Heart rate	HR > 100 beats/min
Low ETCO ₂	End-tidal CO ₂	ETCO ₂ < 30 mmHg

Table 10: Endpoint definitions and abnormal thresholds used for window-level labels.

E Model Zoo and References

Table 12 lists all 37 benchmark models, along with their respective spectral tiers, protocol pools, and references or sources; the source URLs for the models are provided. Among these, the three models—DeepSeek-32B, Qwen3-32B, and GLM-4-32B—were implemented via local deployment; the source URLs for their model parameters are

included.

F Detailed Evaluation Protocol

Classification metrics. For each binary task, predicted scores are evaluated using AUROC and AUPRC. Brier score is computed as the mean squared difference between predicted probabilities and binary labels. Fixed-threshold summaries at 0.5 are retained for standard classification outputs where applicable, including F1, precision, sensitivity, specificity, and Brier score. In the posthoc operating-point analysis, thresholds are instead selected on validation predictions to satisfy prespecified specificity targets and then applied unchanged to the held-out test set.

Regression metrics. Continuous targets are evaluated using MAE, RMSE, and R^2 . MAE emphasizes average absolute error, RMSE penalizes larger deviations, and R^2 evaluates explained vari-

Task	Family	Horizon	Samples	Positives	Positive rate
Future hypotension risk	Future risk	5 min	1,287,334	226,598	17.60%
Future hypotension risk	Future risk	10 min	1,266,428	352,782	27.86%
Future hypoxemia risk	Future risk	5 min	2,122,003	52,787	2.49%
Future bradycardia risk	Future risk	5 min	2,029,350	89,328	4.40%
Future tachycardia risk	Future risk	5 min	1,968,380	256,840	13.05%
Future low ETCO ₂ risk	Future risk	5 min	1,932,620	210,191	10.88%
Persistent hypotension	Persistent deterioration	5 min	129,728	51,461	39.67%
Persistent hypotension	Persistent deterioration	10 min	129,600	35,684	27.53%
Persistent bradycardia	Persistent deterioration	5 min	55,736	28,969	51.98%
Persistent tachycardia	Persistent deterioration	5 min	105,072	69,649	66.29%
Persistent low ETCO ₂	Persistent deterioration	5 min	81,234	34,593	42.58%
Minimum future MAP	Regression	5 min	1,461,123	–	–
Minimum future SpO ₂	Regression	5 min	2,136,422	–	–

Table 11: Canonical window-level tasks and final sample counts. Positive counts are shown for binary classification tasks.

ation relative to the empirical target distribution. For MAP and SpO₂ regression tasks, additional abnormal-region analyses evaluate errors in clinically low-value regions.

Clustered uncertainty. For canonical posthoc prediction files, bootstrap confidence intervals are computed by resampling the exported group identifier rather than individual rows. The group identifier is derived from patient or case identifiers during prediction export, preserving within-patient or within-case dependence among correlated windows. Auxiliary JSONL posthoc utilities may use row-level bootstrap and are not used to define the clustered benchmark uncertainty protocol.

Calibration. Calibration summaries for classification tasks include Brier score, expected calibration error (ECE), and calibration intercept and slope when the fitted calibration model is available. ECE is computed from quantile-based calibration bins, with duplicate bins dropped when score variation is limited. Bin-level calibration summaries are preserved in machine-readable posthoc outputs, and the delivered posthoc figure utilities can render calibration plots from these bin-level files.

Operating points and alert burden. For classification tasks with validation predictions, thresholds are selected on the validation split to target specificity levels of 0.90 and 0.95, then applied to the held-out test split. The resulting operating-point summaries include sensitivity, specificity,

PPV, NPV, and confusion-count components. For window-level warning tasks with timing metadata, alert windows are collapsed into alert episodes using a 300-second refractory period and normalized by eligible monitoring hours and by case, yielding alert episodes, false alert episodes, alerts per hour, false alerts per hour, and alerts per case.

Abnormal-region regression. For continuous MAP and SpO₂ targets, abnormal-region analyses stratify errors by clinically relevant thresholds. MAP is evaluated at 65, 60, and 55 mmHg, and SpO₂ at 92%, 90%, and 85%. For each threshold, the analysis reports overall and threshold-stratified error summaries, miss rate in the true-abnormal region, and threshold-event discrimination using AUROC and AUPRC. This identifies models that appear accurate globally but miss low-value regions that matter clinically.

G Complete Protocol-Stratified Ranking Results

The main text reports a compact summary of the benchmark results to highlight the principal performance patterns across protocol-compatible model pools. Since Pool 1 contains 29 supervised or adapted models, listing all models in the main body would substantially reduce readability and make the comparison table overly dense. Therefore, the full protocol-stratified family-level ranking is provided in this appendix.

Table 13 reports the complete model-wise mean

Model	Pool	Reference / source
<i>Numerical / tree-based and feature-based models</i>		
Logistic Regression	P1	(Cox, 1958)
MLP	P1	(Rumelhart et al., 1986)
MiniRocket	P1	(Dempster et al., 2020)
Summary-Feature MLP	P1	(Johnson et al., 2022)
Random Forest	P1	(Breiman, 2001)
XGBoost	P1	(Chen and Guestrin, 2016)
LightGBM	P1	(Ke et al., 2017)
CatBoost	P1	(Prokhorenkova et al., 2018)
<i>Deep sequential neural networks</i>		
CNN1D	P1	(Wang et al., 2017)
ResNet1D	P1	(He et al., 2016)
InceptionTime	P1	(Ismail Fawaz et al., 2020)
TCN	P1	(Bai et al., 2018)
LSTM	P1	(Hochreiter and Schmidhuber, 1997)
GRU	P1	(Cho et al., 2014)
Transformer Encoder	P1	(Vaswani et al., 2017)
<i>Forecasting and time-series foundation-style models</i>		
Autoformer	P1	(Wu et al., 2021)
FEDformer	P1	(Zhou et al., 2022)
DLinear	P1	(Zeng et al., 2023)
PatchTST	P1	(Nie et al., 2023)
iTransformer	P1	(Liu et al., 2024)
TimesNet	P1	(Wu et al., 2023)
Transformer-TS	P1	(Vaswani et al., 2017)
TSMixer	P1	(Chen et al., 2023)
TFT	P1	(Lim et al., 2021)
N-BEATS	P1	(Oreshkin et al., 2020)
N-HiTS	P1	(Challu et al., 2023)
MOMENT	P1	(Goswami et al., 2024)
Chronos	P2	(Ansari et al., 2024)
TimesFM	P2	(Das et al., 2024)
Moirai	P2	(Woo et al., 2024)
Lag-LLaMA	P2	(Rasul et al., 2024)
Granite-TTM	P2	(Ekambaram et al., 2024)
<i>LLM-based and language-interface endpoints</i>		
TimeLLM	P1	(Jin et al., 2024)
GPT4MTS	P1	(Jia et al., 2024)
DeepSeek-32B	P3	DeepSeek-R1-Distill-Qwen-32B model card
Qwen3-32B	P3	Qwen3-32B model card
GLM-4-32B	P3	GLM-4-32B-0414 model card

Table 12: Model zoo, protocol pools, and model-source references. P1, P2, and P3 denote Pool 1, Pool 2, and Pool 3, respectively.

ranks across task families. Lower ranks indicate better performance. Importantly, ranks are computed only within protocol-compatible pools, rather than across all models globally. Thus, Pool 1, Pool 2, and Pool 3 should be interpreted as separate comparison groups with different training, adaptation, or inference assumptions.

H Extended Results

H.1 Mean-Rank Aggregation

To summarize performance across heterogeneous endpoints, we compute mean ranks rather than averaging raw metric values. Ranking is performed

separately within each protocol-compatible pool, so models from Pool 1, Pool 2, and Pool 3 are not merged into a single global ranking set.

For each task t and pool p , all models with a valid test-set result for that task are ranked according to the task’s primary metric. For binary classification tasks, the primary metric is test AUROC, with larger values ranked better. For the Pool 1 and Pool 2 regression summaries generated by the benchmark code, the primary metric is test R^2 , also ranked with larger values better. If an external regression table is summarized using an error metric such as RMSE, the ranking direction is reversed

Model	Paradigm	Future risk	Persistent det.	Vital-sign reg.	Case-level prog.	Overall rank
Pool 1: supervised / adapted full-test ranking						
CatBoost	Tree-based	1.333	<u>2.000</u>	6.500	1.400	2.808
ResNet1D	Deep sequential	4.000	6.400	2.500	–	<u>4.300</u>
Random Forest	Tree-based	5.333	1.400	10.000	<u>2.600</u>	4.833
LightGBM	Tree-based	4.000	3.600	9.500	3.400	5.125
XGBoost	Tree-based	<u>2.667</u>	3.800	13.000	<u>2.600</u>	5.517
GRU	Deep sequential	8.333	6.200	<u>4.500</u>	–	6.344
InceptionTime	Deep sequential	5.667	9.200	<u>4.500</u>	–	6.456
CNN1D	Deep sequential	8.333	9.000	5.500	–	7.611
LSTM	Deep sequential	8.833	11.600	6.000	–	8.811
Transformer Encoder	Deep sequential	11.667	9.600	8.000	–	9.756
TCN	Deep sequential	13.500	17.800	9.000	–	13.433
N-BEATS	Forecasting / TS	14.500	16.400	11.000	–	13.967
TSMixer	Forecasting / TS	14.833	19.600	11.500	–	15.311
GPT4MTS	LLM-oriented adapted	11.167	8.600	28.000	–	15.922
PatchTST	Forecasting / TS	15.333	14.000	19.500	–	16.278
TimesNet	Forecasting / TS	17.167	13.000	19.000	–	16.389
N-HiTS	Forecasting / TS	17.667	16.800	19.500	–	17.989
iTransformer	Forecasting / TS	18.667	14.400	21.000	–	18.022
Logistic Regression	Linear reference	27.167	27.000	–	5.200	19.789
MiniRocket	Time-series transform	23.167	24.000	12.500	–	19.889
Summary-Feature MLP	Summary neural	27.333	27.000	19.500	6.200	20.008
MLP	Neural reference	27.167	27.000	19.500	6.400	20.017
TimeLLM	LLM-oriented adapted	19.500	18.000	23.000	–	20.167
FEDformer	Forecasting / TS	20.667	21.000	19.000	–	20.222
Transformer-TS	Forecasting / TS	22.667	22.800	15.500	–	20.322
TFT	Forecasting / TS	22.167	15.400	23.500	–	20.356
DLinear	Forecasting / TS	17.333	22.600	22.000	–	20.644
Autoformer	Forecasting / TS	20.333	21.800	22.000	–	21.378
MOMENT	TS foundation / adapted	22.167	22.000	21.000	–	21.722
Pool 2: direct pretrained-transfer full-test comparison						
Chronos	TS foundation / direct	1.500	2.600	1.000	–	1.700
TimesFM	TS foundation / direct	2.667	<u>2.400</u>	<u>2.000</u>	–	<u>2.356</u>
Moirai	TS foundation / direct	<u>1.833</u>	1.800	4.000	–	2.544
Granite-TTM	TS foundation / direct	4.000	3.800	3.000	–	3.600
Lag-LLaMA	TS foundation / direct	5.000	4.400	5.000	–	4.800
Pool 3: Prompt-based LLM language-interface comparison						
<i>With temporal-window input</i>						
DeepSeek-32B	LLM interface + TS	3.000	<u>1.800</u>	1.500	2.800	2.275
Qwen3-32B	LLM interface + TS	<u>1.667</u>	1.600	<u>2.000</u>	<u>2.000</u>	1.817
GLM-4-32B	LLM interface + TS	1.333	2.600	2.500	1.200	<u>1.908</u>
<i>no-temporal-window</i>						
DeepSeek-32B	LLM interface	2.500	2.400	3.000	3.000	2.725
Qwen3-32B	LLM interface	<u>2.000</u>	1.400	<u>2.000</u>	<u>2.000</u>	<u>1.850</u>
GLM-4-32B	LLM interface	1.500	<u>2.200</u>	1.000	1.000	1.425

Table 13: Complete protocol-stratified summary of benchmark performance across task families. Entries report family-level mean ranks, where lower is better. Ranks are computed within protocol-compatible evaluation pools and are not merged across Pool 1, Pool 2, and Pool 3. For Pool 3, prompt-based LLM results are reported separately for prompts with temporal-window input and without temporal-window input. Bold values indicate the best rank within each protocol pool or Pool 3 input setting, and underlined values indicate the second-best rank. Missing entries indicate that the model was not evaluated for that task family under the corresponding protocol.

so that smaller error receives the better rank. Ties are assigned the best shared rank, and subsequent ranks are skipped.

Let $r_{m,t}^{(p)}$ denote the rank of model m on task t within pool p . Tasks are grouped into four task families: future-risk classification, persistent-deterioration classification, continuous vital-sign regression, and case-level prognosis. The family-level mean rank for model m in family g is

$$\bar{r}_{m,g}^{(p)} = \frac{1}{|\mathcal{T}_{m,g}^{(p)}|} \sum_{t \in \mathcal{T}_{m,g}^{(p)}} r_{m,t}^{(p)},$$

where $\mathcal{T}_{m,g}^{(p)}$ is the set of tasks in family g for which model m has an eligible result in pool p . Thus, the family-level number gives each endpoint within a family equal weight.

The overall mean rank is then computed as the unweighted mean of the available family-level

mean ranks:

$$\bar{r}_{m,\text{overall}}^{(p)} = \frac{1}{|\mathcal{G}_m^{(p)}|} \sum_{g \in \mathcal{G}_m^{(p)}} \bar{r}_{m,g}^{(p)},$$

where $\mathcal{G}_m^{(p)}$ is the set of task families with at least one eligible result for model m in pool p . This second-stage averaging gives the four task families equal influence when they are available, rather than allowing families with more endpoints to dominate the summary. Models with missing task families are summarized over their available families only, and the number of contributing families is reported alongside the overall rank.

H.2 Detailed Pool 1 Supervised and Adapted Results

This appendix provides the detailed supervised/adapted Pool 1 results supporting Section 5. Table 14 lists the full Pool 1 family-level mean-rank summary. Tables 15 and 16 report task-level classification and regression leaders. Missing entries indicate that the model was not evaluated for that task family under the Pool 1 protocol.

H.3 Detailed Pool 2 Direct Pretrained-Transfer Results

Table 17 reports the Pool 2 family-level ranks for Chronos, TimesFM, Moirai, Lag-LLaMA, and Granite-TTM. Table 18 gives task-level classification winners within Pool 2, and Table 19 lists the available direct-transfer regression metrics. These rows are direct pretrained-transfer comparisons and are not mixed with Pool 1 ranking.

H.4 Detailed Pool 3 LLM Language-Interface Results

This appendix gives the detailed Pool 3 prompt-based LLM language-interface results. Tables 20 and 21 separate temporal-window and no-temporal-window inputs. The available LLM result files contain 54 rows for temporal-window input, 54 rows for no-temporal-window input, and 108 posthoc bootstrap-summary rows. No separate output-parsing-failure table or missing-output count is present in the available LLM result files; parsing-failure rates are therefore not reported.

LLM prompt construction and output parsing. For each prediction task, we used a structured prompt that explicitly specified the task name, task objective, current-window textual context, and

case-level profile information. In the temporal-window setting, the prompt additionally included the raw 60-second, 1 Hz vital-sign window with four channels: heart rate (HR), mean arterial pressure (MAP), peripheral oxygen saturation (SpO₂), and end-tidal CO₂ (ETCO₂). Each channel was represented as a sequence of 60 values, with missing observations denoted as NA. In the no-temporal-window setting, this raw vital-sign block was removed while all other prompt components were kept unchanged, enabling a controlled comparison between text-only and text-plus-temporal-window LLM inputs. Finally, if the current data retrieval fails or the resulting JSON is invalid, a retry mechanism is triggered.

H.5 Extended Clinical Reliability Analyses

This appendix provides extended reliability-oriented analyses for calibration, validation-selected operating points, alert burden, threshold-event regression, abnormal-region regression, and clustered bootstrap confidence intervals. These analyses complement discrimination and global regression metrics rather than replacing them.

I Code Release and Reproducibility Package

To support reproducible use of PERIBENCH, the released codebase will be organized around the complete benchmark lifecycle: data reconstruction, task generation, model evaluation, reliability analysis, language-interface evaluation, and leaderboard submission. Given authorized access to VitalDB, users will be able to reconstruct the standardized 60-second physiological windows, physiological quality-control masks, window-level tensor manifests, task labels, regression targets, and subject-grouped train/validation/test splits used in the benchmark. The release will also include model configuration files, random seeds, validation criteria, checkpoint references, prediction files, and evaluation scripts for recomputing all reported primary metrics, protocol-specific rankings, and posthoc reliability analyses. For Pool 3 language-interface experiments, the package will provide prompt templates, temporal-window serialization code, case-profile formatting code, response parsers, probability and value conversion rules, retry handling, and failure logs. When raw VitalDB records cannot be redistributed due to source-dataset licensing, the released scripts, mani-

Model	Paradigm	Future risk	Persistent det.	Vital-sign reg.	Case-level prog.	Overall rank
CatBoost	Tree-based	1.333	2.000	6.500	1.400	2.808
ResNet1D	Deep sequential	4.000	6.400	2.500	–	4.300
Random Forest	Tree-based	5.333	1.400	10.000	2.600	4.833
LightGBM	Tree-based	4.000	3.600	9.500	3.400	5.125
XGBoost	Tree-based	2.667	3.800	13.000	2.600	5.517
GRU	Deep sequential	8.333	6.200	4.500	–	6.344
InceptionTime	Deep sequential	5.667	9.200	4.500	–	6.456
CNN1D	Deep sequential	8.333	9.000	5.500	–	7.611
LSTM	Deep sequential	8.833	11.600	6.000	–	8.811
Transformer Encoder	Deep sequential	11.667	9.600	8.000	–	9.756
TCN	Deep sequential	13.500	17.800	9.000	–	13.433
N-BEATS	Forecasting / TS	14.500	16.400	11.000	–	13.967
TSMixer	Forecasting / TS	14.833	19.600	11.500	–	15.311
GPT4MTS	LLM-oriented adapted	11.167	8.600	28.000	–	15.922
PatchTST	Forecasting / TS	15.333	14.000	19.500	–	16.278
TimesNet	Forecasting / TS	17.167	13.000	19.000	–	16.389
N-HiTS	Forecasting / TS	17.667	16.800	19.500	–	17.989
iTransformer	Forecasting / TS	18.667	14.400	21.000	–	18.022
Logistic Regression	Linear reference	27.167	27.000	–	5.200	19.789
MiniRocket	Time-series transform	23.167	24.000	12.500	–	19.889
Summary-Feature MLP	Summary neural	27.333	27.000	19.500	6.200	20.008
MLP	Neural reference	27.167	27.000	19.500	6.400	20.017
TimeLLM	LLM-oriented adapted	19.500	18.000	23.000	–	20.167
FEDformer	Forecasting / TS	20.667	21.000	19.000	–	20.222
Transformer-TS	Forecasting / TS	22.667	22.800	15.500	–	20.322
TFT	Forecasting / TS	22.167	15.400	23.500	–	20.356
DLinear	Forecasting / TS	17.333	22.600	22.000	–	20.644
Autoformer	Forecasting / TS	20.333	21.800	22.000	–	21.378
MOMENT	TS foundation / adapted	22.167	22.000	21.000	–	21.722

Table 14: Full Pool 1 family-level mean ranks. Lower is better. Values are copied from the protocol-aware ranking table and its full Pool 1 continuation.

Family	Task	Best AUROC	Best AUPRC	Test n
Future risk	Future bradycardia, 5 min	CatBoost, 0.915	CatBoost, 0.485	391,642
Future risk	Future hypotension, 5 min	XGBoost, 0.801	CatBoost, 0.549	248,626
Future risk	Future hypotension, 10 min	CatBoost, 0.756	CatBoost, 0.595	244,531
Future risk	Future hypoxemia, 5 min	ResNet1D, 0.808	ResNet1D, 0.267	409,536
Future risk	Future low ETCO ₂ , 5 min	CatBoost, 0.822	CatBoost, 0.463	373,227
Future risk	Future tachycardia, 5 min	CatBoost, 0.818	CatBoost, 0.494	379,961
Persistent deterioration	Persistent bradycardia, 5 min	Random Forest, 0.730	Random Forest, 0.734	10,773
Persistent deterioration	Persistent hypotension, 5 min	Random Forest, 0.754	Random Forest, 0.689	25,065
Persistent deterioration	Persistent hypotension, 10 min	Random Forest, 0.792	Random Forest, 0.644	25,016
Persistent deterioration	Persistent low ETCO ₂ , 5 min	CatBoost, 0.823	CatBoost, 0.786	15,681
Persistent deterioration	Persistent tachycardia, 5 min	CatBoost, 0.852	CatBoost, 0.920	20,295
Case-level prognosis	In-hospital death	CatBoost, 0.834	XGBoost, 0.103	1,230
Case-level prognosis	ICU admission	CatBoost, 0.892	CatBoost, 0.717	1,230
Case-level prognosis	Estimated blood loss ≥ 500 mL	CatBoost, 0.796	Random Forest, 0.553	768
Case-level prognosis	Estimated blood loss ≥ 1000 mL	CatBoost, 0.873	CatBoost, 0.504	768
Case-level prognosis	Any transfusion	Random Forest, 0.817	CatBoost, 0.356	1,229

Table 15: Pool 1 task-level classification winners. Each winner is selected within the supervised/adapted full-test ranking pool.

Task	Best MAE	Best RMSE	Best R^2	Label-available n
Minimum MAP, next 5 min	CNN1D, 6.432	CatBoost, 10.397	CatBoost, 0.568	282,399
Minimum SpO ₂ , next 5 min	iTransformer, 0.924	ResNet1D, 2.374	ResNet1D, 0.425	412,669

Table 16: Pool 1 regression metric leaders. Lower is better for MAE and RMSE; higher is better for R^2 . The final column reports label-available test windows; the iTransformer minimum-SpO₂ MAE was computed on 408,275 windows satisfying its contiguous-window eligibility requirement.

Model	Future risk	Persistent det.	Vital-sign reg.	Case-level prog.	Overall rank
Chronos	1.500	2.600	1.000	–	1.700
TimesFM	2.667	2.400	2.000	–	2.356
Moirai	1.833	1.800	4.000	–	2.544
Granite-TTM	4.000	3.800	3.000	–	3.600
Lag-LLaMA	5.000	4.400	5.000	–	4.800

Table 17: Pool 2 direct pretrained-transfer mean ranks. Lower is better.

Family	Task	Best AUROC	Best AUPRC	Test n
Future risk	Future bradycardia, 5 min	Chronos, 0.888	Chronos, 0.420	387,650
Future risk	Future hypotension, 5 min	Chronos, 0.766	Granite-TTM, 0.488	246,533
Future risk	Future hypotension, 10 min	Chronos, 0.713	Granite-TTM, 0.538	241,262
Future risk	Future hypoxemia, 5 min	Moirai, 0.735	Chronos, 0.157	404,855
Future risk	Future low ETCO ₂ , 5 min	Moirai, 0.747	Granite-TTM, 0.369	371,170
Future risk	Future tachycardia, 5 min	Moirai, 0.779	Moirai, 0.396	376,669
Persistent deterioration	Persistent bradycardia, 5 min	Moirai, 0.689	Moirai, 0.677	10,642
Persistent deterioration	Persistent hypotension, 5 min	Moirai, 0.676	Moirai, 0.561	24,975
Persistent deterioration	Persistent hypotension, 10 min	Moirai, 0.711	Moirai, 0.503	24,743
Persistent deterioration	Persistent low ETCO ₂ , 5 min	Moirai, 0.698	Moirai, 0.589	15,433
Persistent deterioration	Persistent tachycardia, 5 min	Chronos, 0.770	Chronos, 0.846	20,018

Table 18: Pool 2 task-level classification winners. Winners are selected only among direct pretrained-transfer models.

Task	Model	MAE	RMSE	R^2
Minimum MAP, next 5 min	Chronos	8.243	13.749	0.238
Minimum MAP, next 5 min	TimesFM	8.537	14.248	0.182
Minimum MAP, next 5 min	Moirai	9.020	15.659	0.012
Minimum MAP, next 5 min	Granite-TTM	9.399	15.652	0.012
Minimum MAP, next 5 min	Lag-LLaMA	72.584	406.155	-664.022
Minimum SpO ₂ , next 5 min	Chronos	0.794	2.992	-0.048
Minimum SpO ₂ , next 5 min	TimesFM	0.820	3.031	-0.076
Minimum SpO ₂ , next 5 min	Granite-TTM	0.908	3.568	-0.490
Minimum SpO ₂ , next 5 min	Moirai	1.201	6.304	-3.651
Minimum SpO ₂ , next 5 min	Lag-LLaMA	9.695	144.010	-2426.155

Table 19: Full Pool 2 regression metrics. Lower is better for MAE and RMSE; higher is better for R^2 .

fests, and task registry will enable local benchmark reconstruction by users with authorized source-data access. Table 29 summarizes the planned reproducibility artifacts.

Input setting	Model	Task family	Mean AUROC	Mean AUPRC	Mean Brier
With temporal-window input	DeepSeek-32B	Future risk	0.573	0.177	0.122
With temporal-window input	Qwen3-32B	Future risk	0.648	0.194	0.176
With temporal-window input	GLM-4-32B	Future risk	0.693	0.246	0.115
With temporal-window input	DeepSeek-32B	Persistent deterioration	0.577	0.516	0.412
With temporal-window input	Qwen3-32B	Persistent deterioration	0.607	0.522	0.407
With temporal-window input	GLM-4-32B	Persistent deterioration	0.578	0.508	0.370
With temporal-window input	DeepSeek-32B	Case-level prognosis	0.677	0.302	0.088
With temporal-window input	Qwen3-32B	Case-level prognosis	0.722	0.253	0.098
With temporal-window input	GLM-4-32B	Case-level prognosis	0.797	0.291	0.253
no-temporal-window	DeepSeek-32B	Future risk	0.541	0.146	0.111
no-temporal-window	Qwen3-32B	Future risk	0.588	0.163	0.226
no-temporal-window	GLM-4-32B	Future risk	0.660	0.204	0.138
no-temporal-window	DeepSeek-32B	Persistent deterioration	0.577	0.519	0.374
no-temporal-window	Qwen3-32B	Persistent deterioration	0.616	0.533	0.349
no-temporal-window	GLM-4-32B	Persistent deterioration	0.567	0.504	0.346
no-temporal-window	DeepSeek-32B	Case-level prognosis	0.658	0.291	0.091
no-temporal-window	Qwen3-32B	Case-level prognosis	0.728	0.267	0.097
no-temporal-window	GLM-4-32B	Case-level prognosis	0.797	0.293	0.253

Table 20: Pool 3 family-level classification metrics for prompt-based LLMs. Means are computed over tasks in each family within the corresponding input setting.

Input setting	Model	Task	MAE	RMSE	R^2
With temporal-window input	DeepSeek-32B	Minimum MAP, next 5 min	7.725	13.913	0.190
With temporal-window input	Qwen3-32B	Minimum MAP, next 5 min	7.166	13.140	0.277
With temporal-window input	GLM-4-32B	Minimum MAP, next 5 min	7.848	14.069	0.171
With temporal-window input	DeepSeek-32B	Minimum SpO ₂ , next 5 min	0.773	2.637	0.394
With temporal-window input	Qwen3-32B	Minimum SpO ₂ , next 5 min	0.801	2.668	0.380
With temporal-window input	GLM-4-32B	Minimum SpO ₂ , next 5 min	0.775	2.657	0.385
no-temporal-window	DeepSeek-32B	Minimum MAP, next 5 min	10.086	17.982	-0.354
no-temporal-window	Qwen3-32B	Minimum MAP, next 5 min	8.551	15.794	-0.044
no-temporal-window	GLM-4-32B	Minimum MAP, next 5 min	8.249	15.419	0.005
no-temporal-window	DeepSeek-32B	Minimum SpO ₂ , next 5 min	3.594	4.473	-0.744
no-temporal-window	Qwen3-32B	Minimum SpO ₂ , next 5 min	1.750	3.055	0.187
no-temporal-window	GLM-4-32B	Minimum SpO ₂ , next 5 min	1.329	3.024	0.203

Table 21: Pool 3 regression metrics for prompt-based LLMs with temporal-window input and without temporal-window input.

Task type	System prompt	Input fields	Required output
Binary classification with time-series window	You are a binary classification model for predicting perioperative risk. Using only the text and numerical information from the current window, as well as the patient's case-profile information, return valid JSON—for example: a JSON object with fields "label" and "probability", e.g., {"label": 1, "probability": 0.73}. Here, "label" represents the class label, and "probability" represents the probability.	system prompt + window_context_text + case_profile_text + raw_1 Hz_60s_channels: HR = [...60 values...] MAP = [...60 values...] SpO ₂ = [...60 values...] ETCO ₂ = [...60 values...]	{"label": 0 or 1, "probability": number from 0 to 1}.
Binary classification without time-series window	Same as above.	system prompt + window_context_text + case_profile_text	{"label": 0 or 1, "probability": number from 0 to 1}.
Continuous regression with time-series window	You are a regression model for predicting perioperative risk. Using only the text and numerical information within the current window, as well as the patient's case notes, return fully valid JSON data—for example: a JSON object with a numeric "value" field, e.g., {"value": 72.3}, where "value" represents your predicted numerical output.	system prompt + window_context_text + case_profile_text + raw_1 Hz_60s_channels: HR = [...60 values...] MAP = [...60 values...] SpO ₂ = [...60 values...] ETCO ₂ = [...60 values...]	{"value": number}.
Continuous regression without time-series window	Same as above.	system prompt + window_context_text + case_profile_text	{"value": number}.

Table 22: Prompt templates and output parsing rules for LLM-based prediction tasks.

Task	Model	AUROC	AUPRC	Brier	ECE
Future bradycardia, 5 min	CatBoost	0.915	0.485	0.030	0.005
Future hypoxemia, 5 min	ResNet1D	0.808	0.267	0.021	0.004
Persistent hypotension, 5 min	Random Forest	0.754	0.689	0.195	0.057
Persistent tachycardia, 5 min	CatBoost	0.852	0.920	0.147	0.020
ICU admission	CatBoost	0.892	0.717	0.090	0.023

Table 23: Representative calibration and probability-error results. ECE is reported when calibration bins are available.

Task	Model	Target spec.	Threshold	Sens.	Spec.	PPV	False alerts/hr	Alerts/case
Future hypoxemia, 5 min	ResNet1D	0.95	0.086	0.484	0.955	0.214	0.884	4.958
Persistent tachycardia, 5 min	CatBoost	0.95	0.831	0.534	0.938	0.944	0.227	1.071
Persistent hypotension, 5 min	Random Forest	0.95	0.689	0.332	0.938	0.779	0.308	1.169

Table 24: Validation-selected operating points and alert burden at a target specificity of 0.95. Alert burden uses a 300-second refractory period and merged eligible monitoring intervals by case.

Threshold-event task	Best AUROC	Best AUPRC
Minimum MAP, next 5 min, threshold 65	LightGBM, 0.878	ResNet1D, 0.816
Minimum SpO ₂ , next 5 min, threshold 90	CatBoost, 0.840	ResNet1D, 0.356

Table 25: Threshold-event analyses for continuous regression targets.

Task	Threshold	Region	Model	<i>n</i>	MAE	RMSE	Miss rate
Minimum MAP, next 5 min	65	overall	CatBoost	282,399	6.584	10.397	–
Minimum MAP, next 5 min	65	true_below_65	CatBoost	76,463	10.403	16.815	0.270
Minimum MAP, next 5 min	65	true_below_65	N-HiTS	76,463	9.734	14.978	0.187
Minimum SpO ₂ , next 5 min	90	overall	CatBoost	408,275	1.296	2.838	–
Minimum SpO ₂ , next 5 min	90	true_below_90	CatBoost	6,290	10.579	13.409	0.684
Minimum SpO ₂ , next 5 min	90	true_below_90	MOMENT	6,290	6.610	11.703	0.000

Table 26: Representative abnormal-region regression results. Minimum-SpO₂ results use a shared set of 408,275 eligible test windows, including 6,290 true-below-90 cases for both models.

Task	Model	Metric	Point	Clustered bootstrap 95% CI
Future bradycardia, 5 min	CatBoost	AUROC	0.915	[0.906, 0.924]
Future bradycardia, 5 min	CatBoost	AUPRC	0.485	[0.453, 0.516]
Future bradycardia, 5 min	CatBoost	BRIER	0.030	[0.027, 0.033]
Future hypoxemia, 5 min	ResNet1D	AUROC	0.808	[0.790, 0.824]
Future hypoxemia, 5 min	ResNet1D	AUPRC	0.267	[0.232, 0.302]
Future hypoxemia, 5 min	ResNet1D	BRIER	0.021	[0.018, 0.023]
Persistent hypotension, 5 min	Random Forest	AUROC	0.754	[0.716, 0.787]
Persistent hypotension, 5 min	Random Forest	AUPRC	0.689	[0.603, 0.751]
Persistent hypotension, 5 min	Random Forest	BRIER	0.195	[0.184, 0.205]
ICU admission	CatBoost	AUROC	0.892	[0.869, 0.917]
ICU admission	CatBoost	AUPRC	0.717	[0.655, 0.776]
ICU admission	CatBoost	BRIER	0.090	[0.078, 0.102]

Table 27: Representative clustered bootstrap confidence intervals for classification metrics. Each row uses 1000 bootstrap resamples.

Task	Threshold	Region	Model	Metric	Point	Clustered bootstrap 95% CI
Minimum MAP, next 5 min	65	true_below_65	N-HiTS	MAE	9.734	[9.114, 10.392]
Minimum MAP, next 5 min	65	true_below_65	N-HiTS	RMSE	14.978	[14.167, 15.838]
Minimum MAP, next 5 min	65	true_below_65	N-HiTS	MISS_RATE	0.187	[0.168, 0.209]
Minimum SpO ₂ , next 5 min	90	true_below_90	MOMENT	MAE	6.610	[5.816, 7.411]
Minimum SpO ₂ , next 5 min	90	true_below_90	MOMENT	RMSE	11.703	[10.549, 13.096]
Minimum SpO ₂ , next 5 min	90	true_below_90	MOMENT	MISS_RATE	0.000	[0.000, 0.000]

Table 28: Representative clustered bootstrap confidence intervals for abnormal-region regression metrics. Each row uses 1000 bootstrap resamples.

Artifact	Release content	Purpose
Preprocessing code	VitalDB loading, physiological quality control, 1 Hz resampling, 60-second window construction, tensor export, and validity-mask generation.	Reconstruct standardized physiological inputs from authorized source records.
Task registry	Canonical definitions of task families, prediction horizons, endpoint thresholds, eligibility filters, validity requirements, label rules, and regression targets.	Rebuild task labels and targets under the same benchmark semantics.
Split manifests	Subject-grouped train/validation/test identifiers and window-level manifests containing case, subject, timestamp, task, and split fields.	Reproduce leakage-controlled partitions and task-specific eligible samples.
Model configurations	Hyperparameters, random seeds, validation criteria, checkpoint references, normalization settings, and training/evaluation commands.	Reproduce model fitting, checkpoint selection, and full-test evaluation.
Prediction files	Per-task prediction outputs for all reported systems, including scores, probabilities, regression values, and eligible test identifiers.	Recompute metrics and posthoc analyses without retraining all models.
Evaluation code	Scripts for AUROC, AUPRC, Brier score, F1, precision, sensitivity, specificity, MAE, RMSE, R^2 , and family-level rank aggregation.	Reproduce the main benchmark tables and protocol-specific rankings.
Reliability analysis	Calibration, validation-selected operating points, alert burden, abnormal-region error analysis, threshold-event analysis, and clustered bootstrap utilities.	Recompute deployment-oriented and clinical posthoc analyses.
Language-interface package	Prompt templates, temporal-window serialization, case-profile formatting, response parsing, probability/value conversion, retry handling, and failure logs.	Reproduce Pool 3 prompt-based LLM interface evaluation.
Leaderboard specification	Task list, allowed inputs, eligible splits, metric direction, submission schema, validation checks, and reporting rules.	Standardize external submissions under protocol-compatible comparison.

Table 29: Planned reproducibility artifacts. The release separates source-data reconstruction, benchmark metadata, model outputs, evaluation scripts, reliability-analysis utilities, language-interface components, and leaderboard rules, enabling users with authorized VitalDB access to rebuild the benchmark and recompute the reported results.