Evaluating Language Models as Descriptors of Neonatal Heart Rate in Mortality Prediction

Medhasweta Sen
School of Data Science
University of Virginia
jwm9fu@virginia.edu

Jiaxing Qiu
School of Data Science
University of Virginia

Tom Hartvigsen
School of Data Science
University of Virginia

Abstract

In Neonatal Intensive Care Units (NICUs) heart-rate monitoring produces continuous time-series signals that, combined with clinical metadata, are critical for early warning and decision support. Traditional statistical models cannot not effectively incorporate textual inputs, leaving clinical information unused in prediction. Recent advances in multimodal language models (LMs) enable aligning temporal signals with textual clinical metadata. We propose a two-step framework to test whether combining numerical time-series and clinical text yields better predictions, by: first, testing LMs' recognition and differentiation capabilities of clinical descriptions tied to temporal and visual properties of NICU heart rate signals; second, evaluating the transfer of this ability to a downstream clinically significant task of 7-day mortality prediction . Results show that descriptive performance strongly correlates to mortality prediction accuracy, with patient metadata and clinical descriptions boosting outcomes, especially for larger models. Vision-Language Models (VLMs) perform best overall, while specialized Time Series Language Models (TSLMs) consistently surpass their base large language models (LLMs). Overall our work provides (1) a controlled evaluation framework linking time series understanding to clinically meaningful downstream tasks, (2) quantification of the added value of metadata and descriptions, and (3) evidence that aligning time series with linguistic understanding is transferable to high-stakes clinical tasks.

1 Introduction

Continuous physiological monitoring in neonatal intensive care units (NICUs) produces rich streams of time-series signals that, when combined with patient history, are critical for early warning systems and clinical decision support [1, 2, 3, 4]. Traditionally, most work with NICU data has relied on numerical features processed through conventional machine learning or statistical models [5, 6, 7, 8, 9, 10]. These approaches, however, have largely ignored the role of textual inputs because such methods could not directly incorporate text.

Recent advances in LLMs and, more specifically, TSLMs [11, 12, 13, 14, 15, 16] now make it possible to align temporal signals and patient metadata (e.g., demographics, clinical history) with clinical descriptions (e.g., noting patterns such as "low variability" or "recurrent bradycardia"). This shift introduces new opportunities for text-conditioned time-series analysis. Despite growing interest in LLMs, VLMs and TSLMs, we still lack systematic evidence on whether aligning NICU heart beat time series with interpretable, human-readable clinical descriptions and the addition of patient metadata improves performance on downstream clinical tasks.

We study this question using a large NICU heart-rate dataset from UVA Hospital comprising 36k+ recordings[17], each a sequence of 300 timesteps at 2-second intervals, with associated patient

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Learning from Time Series for Health.

metadata (e.g., gestational age, delivery type, etc.) and natural-language clinical descriptions of time-series morphology (variability, and clinical events such as bradycardia). The 7-day mortality label is highly imbalanced (278 deaths vs. 36,401 survivors), representative of a real world clinical distribution.

Our investigation includes LMs across 3 modalities – LLMs, VLMs and TSLMs and proceeds in two cascading steps:

1. **Experiment 1:** Evaluates LMs' descriptive understanding through two tasks. **Task 1: Recognition.** We prompt models to decide whether a given description is valid for a time series, framed as a True/False classification problem. **Task 2: Differentiation.** We ask models to select the correct description from one true option and three distractors, posed as a multiple-choice problem. This yields a proxy for "clinical time-series understanding."
2. **Experiment 2:** Tests whether the understanding of the time series heart rate data transfers to a clinically meaningful downstream endpoint of 7-day mortality prediction under **six** input-output conditions.

This design lets us test two central hypotheses: **Ranking Consistency:** Model performance on descriptive tasks (recognition and differentiation) is predictive of performance on downstream mortality prediction. **Contextual Enrichment:** Performance improves monotonically as additional context is provided—moving from time series alone to combinations with patient metadata and clinical descriptions.

In this work, we make three main contributions. **First**, we introduce a two-step evaluation: description recognition and differentiation as a proxy for time-series understanding, and a downstream test on real NICU mortality prediction. **Second**, we measure the added value of patient metadata and clinical descriptions—both separately and together—within a controlled experimental framework. **Third**, we show that stronger descriptive understanding in Experiment 1 translates into better mortality prediction in Experiment 2, providing evidence that the ability to align time series with language is a transferable capability for high-stakes clinical tasks.

2 Experimental Set-Up

2.1 Experiments

2.1.1 Experiment 1: Descriptive Understanding

In the first step of experiments, we propose two tasks that capable time-series reasoning models should be able to perform. We evaluate 8 state-of-the-art LMs on their ability to recognize clinical time-series descriptions of NICU patient heart rates under two input conditions: with and without patient metadata. These tasks are formulated as Question–Answering tasks, where models are given time series and asked to recognize or differentiate between correct and incorrect descriptions.

Description Recognition (Task 1) Given a NICU heart rate time series and an accompanying clinical description, a model must determine whether the description is valid. We format this as a True/False task and design the prompt p_i to encourage $f(x_i, d_i) \in \{\text{“True”}, \text{“False”}\}$. Because our datasets do not contain "incorrect" (False) time series and description pairs naturally, we obtain them by negative sampling from within our dataset. We use two methods to ensure the robustness of our benchmark to the choice of sampling method. One method assesses the similarity of captions, while the other assesses the similarity of time series. In all cases, we select the most dissimilar option as the incorrect description while ensuring it is drawn from a time series with the opposite clinical outcome. Appendix 5.2 provides further details about our method for selecting incorrect options.

Description Differentiation (Task 2) Given a time series x_i, the model must select the correct description d_i from a set of four options $\{d_i, d_j, d_k, d_l\}$. This is posed as a multiple-choice task, with the prompt p_i formatted to present the options and elicit a letter-valued prediction $f(x_i, p_i) \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$, in which each letter corresponds to a description option. A prediction is correct if it corresponds to the index of the true description d_i. As in Task 1, incorrect options are generated by sampling the top three most dissimilar descriptions from the dataset, using both negative sampling methods. Note, for both Tasks 1 and 2, we also run parallel experiments in which models are provided with additional patient metadata as context alongside the time series.

2.1.2 Experiment 2: Mortality Prediction

The second step evaluates whether the descriptive understanding measured in Experiment 2 transfers to a downstream task of clinically meaningful endpoint: 7-day mortality prediction. We also test the effect of ground truth clinical descriptions and patient metadata on prediction accuracy. We formulate this as a binary classification task, where models predict whether an infant dies (1) or survives (0). Building on the proxy of time-series understanding from recognition and differentiation, we test how different combinations of input signals (heart rate time series, metadata, and ground truth clinical descriptions) affect downstream predictive performance.

We design six experimental setups:

1. **TS→Pred(E2.1):** Given NICU heart-rate time series alone → predict 7-day mortality.
2. **TS→Desc(E2.2):** Given TS, first generate a brief description, then based on the description predict 7-day mortality.
3. **TS+Desc→Pred(E2.3):** Given TS with a brief ground truth clinical description → predict 7-day mortality.
4. **TS+Meta→Pred(E2.4):** Given TS with patient metadata (e.g., gestational age, delivery type) → predict 7-day mortality.
5. **TS+Meta→Desc→Pred(E2.5):** Given TS and metadata, first generate a brief description, then based on the description predict 7-day mortality.
6. **TS+Meta+Desc→Pred(E2.6):** Given TS with metadata and a brief ground truth clinical description → predict 7-day mortality.

2.2 Models

We evaluate four LLMs, three VLMs, and one time series–language model (TSLM) in both experiments. This includes proprietary models and public models that range from 4.2B to 14B parameters. **LLMs:** GPT-4o [18],Phi-3.5-Mini-Instruct [19], Qwen2.5-14B-Instruct-1M [20], and Qwen2.5-14B-Instruct-1M [20]. We convert time series to strings of comma-separated values, following prior works [21, 22]. **VLMs:** We evaluate GPT-4o-Vision [18], Qwen2.5-VL-7B-Instruct [23], and Phi-3.5-Vision-Instruct [19]. Time series are represented as `matplotlib`-rendered plots. **TSLMs:** We evaluate ChatTS-14B [12], which operates directly on numerical vectors.

2.3 Metrics

In Experiment 1, Recognition and Differentiation are classification tasks with balanced datasets, so we measure Accuracy. For Experiment 2, since the dataset is highly imbalanced, we measure Weighted F1-Score.

3 Results

Models	Recognition			Differentiation			OaR
	TS	TS + Metadata	Rank	TS	TS + Metadata	Rank	
GPT-4o	0.567	0.640	2	0.944	0.971	2	2
GPT-4o-V	0.604	0.667	1	0.957	0.983	1	1
Qwen-14B	0.540	0.612	4	0.814	0.852	3.5	3.75
ChatTS-14B	0.545	0.587	3.5	0.821	0.776	3.5	3.5
Qwen-7B	0.530	0.520	6.5	0.782	0.698	6.5	6.5
Qwen-7B-V	0.544	0.553	4.5	0.793	0.721	5	4.75
Phi-mini	0.506	0.518	8	0.765	0.667	8	8
Phi-mini-V	0.529	0.529	6.5	0.771	0.714	6.5	6.5

Table 1: Recognition & Differentiation Accuracy with DTW-based distractors. OaR = Overall rank

We first compare all models on Recognition and Differentiation tasks. Models that fail to produce outputs in the required format are counted as errors, which can reduce accuracy. To highlight general trends, we also report each model's average rank across both tasks and input settings Table 1. Overall, VLMs consistently outperform their text-only LLM counterparts, with GPT-4o-Vision emerging as the best-performing model across all settings. This advantage is expected, as clinical descriptions depend heavily on visual properties of the time series. Among LLMs, GPT-4o is the strongest, ranking second overall behind its vision-enabled variant. Scaling laws hold: larger models such as Qwen-14B outperform Qwen-7B which surpasses Phi-mini. Notably, ChatTS-14B consistently outperforms its base LLM (Qwen-14B), underscoring the value of architectures tuned specifically to temporal data.

Adding metadata consistently improves Recognition accuracy. However, for Differentiation tasks, metadata provides benefits only to the strongest models, while smaller models often perform worse with additional context. This could be because smaller open source models might not have been exposed to such clinical and demographic information in its' training thus fail to exploit it in a zero-shot setting. Note, all results shown here use distractors selected via DTW distance, which identifies the distractors by finding the most dissimilar time series. The same performance trends hold when distractors are instead selected using Sentence-BERT embeddings with cosine similarity, which identifies the most dissimilar clinical descriptions directly as shown in Table 3.

Models	**E2.1**	**E2.2**	**E2.3**	**E2.4**	**E2.5**	**E2.6**	**Rank**
GPT-4o	0.982	0.952	0.939	0.984	0.943	0.944	1.5
GPT-4o-V	0.983	0.958	0.938	0.986	0.943	0.941	1.5
Qwen-14B	0.660	0.775	0.765	0.742	0.817	0.793	4.5
ChatTS-14B	0.980	0.950	0.909	0.984	0.928	0.914	2.833
Qwen-7B	0.607	0.662	0.648	0.629	0.727	0.743	5.833
Qwen-7B-V	0.432	0.868	0.701	0.733	0.906	0.896	4.667
Phi-mini	0.349	0.464	0.430	0.360	0.480	0.470	8
Phi-mini-V	0.363	0.478	0.431	0.393	0.564	0.619	7

Table 2: Weighted F1 for 7-day mortality prediction across six input settings (E2.1–E2.6)

Experiment 2 evaluates whether descriptive understanding transfers to the clinically meaningful endpoint of 7-day mortality prediction. Results across the six input–output conditions are shown in Table 2. We find the relative ranking of models in mortality prediction is highly correlated their ranking in Experiment 1. Overall, VLMs consistently outperform their LLM counterparts. GPT-4o-Vision remains the top-performing model, followed closely by GPT-4o, confirming that strong descriptive understanding is predictive of downstream clinical performance. Similarly, ChatTS-14B again outperforms its base LLM (Qwen-14B), reinforcing the benefit of temporal specialization observed in Experiment 1. Similarly, scaling laws uphold where in smaller open-source LLMs and VLMs under-perform. This supports our first hypothesis: descriptive tasks serve as a reliable proxy for downstream clinical utility.

Performance generally improves as additional context is provided. Moving from raw time series alone to TS+Patient Metadata or TS+Patient Metadata+Clinical Description yields steady gains. Notably, metadata consistently provides the largest boost to mortality prediction, while descriptions alone provide smaller but still positive improvements, as, the clinical endpoint encourages models to exploit demographic and contextual signals effectively. However, smaller models benefit less from context, consistent with their limited capacity to integrate heterogeneous inputs in a zero-shot setting. Together, these results reveal a positive correlation between descriptive performance in Experiment 1 and improved clinical prediction in Experiment 2. Models that best aligned time series with clinical language also achieved the highest accuracy on mortality, and performance improved monotonically as additional contextual inputs were introduced. This provides strong evidence that the ability to link heart rate temporal signals with interpretable clinical descriptions is not only measurable but also clinically useful.

4 Conclusion

This work introduces a controlled two-step framework linking descriptive understanding of NICU heart-rate time series to clinically meaningful outcomes. Across both experiments, we find that LMs capable of accurately identifying clinical descriptions also achieve stronger performance on 7-day mortality prediction, validating descriptive tasks as a reliable proxy for downstream utility. VLMs consistently lead, while TSLMs outperform their base LLMs, underscoring the value of temporal alignment. Performance further improves with patient metadata and clinical descriptions, confirming the additive benefit of contextual signals. Together, these findings provide evidence that aligning time-series data with language not only advances interpretability but also translates into improved predictions in high-stakes clinical settings.

References

[1] Justin C Niestroy, J Randall Moorman, Maxwell A Levinson, Sadnan Al Manir, Timothy W Clark, Karen D Fairchild, and Douglas E Lake. Discovery of signatures of fatal neonatal illness in vital signs using highly comparative time-series analysis. *NPJ digital medicine*, 5(1):6, 2022.

[2] Jiarui Feng, Jennifer Lee, Zachary A Vesoulis, and Fuhai Li. Predicting mortality risk for preterm infants using deep learning models with time-series vital sign data. *npj Digital Medicine*, 4(1):108, 2021.

[3] Rohan Joshi, Deedee Kommers, Laurien Oosterwijk, Loe Feijs, Carola Van Pul, and Peter Andriessen. Predicting neonatal sepsis using features of heart rate variability, respiratory characteristics, and ecg-derived estimates of infant motion. *IEEE journal of biomedical and health informatics*, 24(3):681–692, 2019.

[4] Hosein Dalili, Mamak Shariat, and Leyla Sahebi. Time series analysis for forecasting neonatal intensive care unit census and neonatal mortality. *BMC pediatrics*, 25(1):339, 2025.

[5] M Pamela Griffin, T Michael O'Shea, Eric A Bissonette, Frank E Harrell, Douglas E Lake, and J Randall Moorman. Abnormal heart rate characteristics are associated with neonatal mortality. *Pediatric research*, 55(5):782–788, 2004.

[6] Navin Kumar, Gangaram Akangire, Brynne Sullivan, Karen Fairchild, and Venkatesh Sampath. Continuous vital sign analysis for predicting and preventing neonatal diseases in the twenty-first century: big data to the forefront. *Pediatric research*, 87(2):210–220, 2020.

[7] Trang Nguyen Phuc Thu, Alfredo I Hernández, Nathalie Costet, Hugues Patural, Vincent Pichot, Guy Carrault, and Alain Beuchée. Improving methodology in heart rate variability analysis for the premature infants: Impact of the time length. *Plos one*, 14(8):e0220692, 2019.

[8] Joel Jaskari, Janne Myllärinen, Markus Leskinen, Ali Bahrami Rad, Jaakko Hollmén, Sture Andersson, and Simo Särkkä. Machine learning methods for neonatal mortality and morbidity classification. *IEEE Access*, 8:123347–123358, 2020.

[9] Joseph Randall Moorman, Waldemar A Carlo, John Kattwinkel, Robert L Schelonka, Peter J Porcelli, Christina T Navarrete, Eduardo Bancalari, Judy L Aschner, Marshall Whit Walker, Jose A Perez, et al. Mortality reduction by heart rate characteristic monitoring in very low birth weight neonates: a randomized trial. *The Journal of pediatrics*, 159(6):900–906, 2011.

[10] Brynne A Sullivan, Christina McClure, Jamie Hicks, Douglas E Lake, J Randall Moorman, and Karen D Fairchild. Early heart rate characteristics predict death and morbidities in preterm infants. *The Journal of pediatrics*, 174:57–62, 2016.

[11] Chengsen Wang, Qi Qi, Jingyu Wang, Haifeng Sun, Zirui Zhuang, Jinming Wu, Lei Zhang, and Jianxin Liao. Chattime: A unified multimodal time series foundation model bridging numerical and textual data. *arXiv preprint arXiv:2412.11376*, 2024.

[12] Zhe Xie, Zeyan Li, Xiao He, Longlong Xu, Xidao Wen, Tieying Zhang, Jianjun Chen, Rui Shi, and Dan Pei. Chatts: Aligning time series with llms via synthetic data for enhanced understanding and reasoning. *arXiv preprint arXiv:2412.03104*, 2024.

[13] Mohamed Trabelsi, Aidan Boyd, Jin Cao, and Huseyin Uzunalioglu. Time series language model for descriptive caption generation, 2025.

[14] Aoi Ito, Kota Dohi, and Yohei Kawaguchi. Clasp: Learning concepts for time-series signals from natural language supervision, 2025.

[15] Winnie Chow, Lauren Gardiner, Haraldur T. Hallgrímsson, Maxwell A. Xu, and Shirley You Ren. Towards time series reasoning with llms, 2024.

[16] Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. *arXiv preprint arXiv:2502.04395*, 2025.

[17] Jiaxing Qiu, Dongliang Guo, Brynne Sullivan, Teague R. Henry, and Tom Hartvigsen. Instruction-based time series editing, 2025.

[18] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

[19] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra,

Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

[20] An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, Weijia Xu, Wenbiao Yin, Wenyuan Yu, Xiafei Qiu, Xingzhang Ren, Xinlong Yang, Yong Li, Zhiying Xu, and Zipeng Zhang. Qwen2.5-1m technical report, 2025.

[21] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. Large language models are zero-shot time series forecasters. In *NeurIPS*, 2023.

[22] Yifu Cai, Arjun Choudhry, Mononito Goswami, and Artur Dubrawski. Timeseriesexam: A time series understanding exam. *arXiv preprint arXiv:2410.14752*, 2024.

[23] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

[24] Ian German Mesner. Pediatric Academic Societies 2024 NICU Mortality Prediction Challenge, 2024.

[25] Brynne A Sullivan, Alvaro G Moreira, Ryan M McAdams, Lindsey A Knake, Ameena Husain, Jiaxing Qiu, Avinash Mudireddy, Abrar Majeedi, Wissam Shalish, Douglas E Lake, et al. Comparing machine learning techniques for neonatal mortality prediction: insights from a modeling competition. *Pediatric research*, pages 1–7, 2024.

[26] Namasivayam Ambalavanan, Debra E Weese-Mayer, Anna Maria Hibbs, Nelson Claure, John L Carroll, J Randall Moorman, Eduardo Bancalari, Aaron Hamvas, Richard J Martin, Juliann M Di Fiore, et al. Cardiorespiratory monitoring data to predict respiratory outcomes in extremely preterm infants. *American journal of respiratory and critical care medicine*, 208(1):79–97, 2023.

5 Appendix

5.1 Dataset Details

We use a publicly dataset of daily heart beat time series observations from 2,964 infants admitted to the University of Virginia NICU between 2012–2016 [24, 25], consisting of 10-minute HR segments (length 300, sampled every 2s). The processed dataset contains 36,679 series, including 2,147 bradycardia events (prevalence 0.06), defined as HR <100 bpm up to 300s [26]. A valid event requires a negative drop rate prior to onset and a positive recovery afterward. Each time series is annotated with one of two event labels — “No events” or “Bradycardia events happened” — and one of two variability labels — “High variability” or “Low variability.” These labels are provided as input to the GPT-4o API, which generates corresponding clinical, human-readable descriptions of the time series.

The patient metadata includes the following variables:

- **EGA** – Estimated gestational age in weeks
- **BWT** – Birth weight in grams
- **Male** – Sex
- **Apgar1** – Apgar 1-minute score
- **Apgar5** – Apgar 5-minute score
- **Vaginal** – Vaginal delivery
- **C-section** – Cesarean delivery
- **Steroids** – Antenatal steroids
- **InBorn** – Born in hospital
- **BirthHC** – Head circumference at birth
- **Multiple** – Multiple births
- **Black, Hispanic, White** – Race
- **MaternalAge** – Maternal age in years

These metadata fields are preprocessed and then passed through the GPT-4o API to generate concrete textual representations of patient history and demographics, which we collectively refer to as patient metadata.

5.2 Selecting Distractors

To support both True/False and Multiple Choice formats in the Recognition and Differentiation tasks, we construct contrastive examples by selecting negative descriptions using four distinct strategies:

- **Caption-based similarity (Sentence-BERT):** We compute cosine similarity over Sentence-BERT embeddings and select descriptions that are semantically dissimilar to the reference.
- **Dynamic Time Warping (DTW):** We measure alignment costs between time series and choose those with the highest DTW distance from the input.

The Sentence-BERT strategy operates over natural language annotations; while the DTW distance directly in time series space. When multiple annotations exist for a given time series, we randomly sample one for evaluation. Negative samples are selected to be maximally dissimilar, simplifying the contrastive setup and providing an upper-bound estimate of model performance. This design ensures that the benchmark evaluates models’ ability to reject clearly incorrect options before advancing to more fine-grained reasoning. Additional check is implemented that all distractor time series has the opposite patient outcome.

5.3 Prompt Details

Experiment 1: Task 1: Only TS

"You are given a neonatal heart-rate time series (bpm, 2s sampling) and a candidate natural-language description. Decide whether the description accurately and adequately reflects salient properties of the series. Answer with ONLY ’YES’ or ’NO’.

Time series: row[’heart rate’]

Clinical Description: row[’desc str’]"

Experiment 1: Task 1: TS+Metadata

"You are given neonatal patient information including demographics, perinatal metadata, and a heart-rate time series (bpm, 2s sampling). Based on the metadata and the time series, decide whether the candidate description accurately and adequately reflects salient properties of the series and patient context such as overall heart rate level, variability, trends, spikes/dips, and clinically relevant context). Answer with ONLY 'YES' or 'NO'.

Time series: row['heart rate']

Metadata:row['patient metadata']

Clinical Description: row['desc str']"

Experiment 1: Task 2: Only TS

"You are given neonatal patient information including demographics, perinatal metadata with neonatal heart-rate time series (bpm, 2s sampling) and four candidate natural-language descriptions. Only ONE description is correct; the other three are incorrect. Choose the option that best represents the time series. Answer ONLY with a single letter A, B, C, or D.

Time series: row['heart rate']

Options:row['options']"

Experiment 1: Task 2: TS+Metadata

"You are given neonatal patient information including demographics, perinatal metadata, with neonatal heart-rate time series (bpm, 2s sampling) and four candidate natural-language descriptions. Only ONE description is correct; the other three are incorrect. Based on the metadata and the time series Choose the option that best represents the time series. Answer ONLY with a single letter A, B, C, or D.

Time series: row['heart rate']

Metadata:row['patient metadata']

Options:row['heart rate']"

Experiment 2.1

"You are given NICU time-series data. Predict whether the infant will die in 7 days,or whether the infant will survive. Respond with **only** a single digit: '1' if the infant will die in 7 days, or '0' if the infant will survive."

Heart rate data: row['heart rate']"

Experiment 2.2

"You are given NICU time-series data. First, generate a brief natural language description of the heart rate pattern you observe. Then, based on that description, predict whether the infant will die in 7 days (**1**) or survive (**0**). "Respond with only the description followed by the single digit decision.

Heart rate data: row['heart rate']"

Experiment 2.3

"You are given NICU time-series data and a brief clinical description of the time series. Predict whether the infant will die in 7 days,or whether the infant will survive. Respond with **only** a single digit: '1' if the infant will die in 7 days, or '0' if the infant will survive.

Heart rate data: row['heart rate']

Clinical description: row['clinical description']"

Experiment 2.4

"You are given NICU time-series data and patient metadata. Predict whether the infant will die in 7 days,or whether the infant will survive. Respond with **only** a single digit: '1' if the infant will die in 7 days, or '0' if the infant will survive.

Heart rate data: row['heart rate']

Patient metadata: row['patient metadata']"

Experiment 2.5

"You are given NICU time-series data and patient metadata. First, generate a brief natural language description of the heart rate pattern you observe. Then, based on that description, predict whether the infant will die in 7 days (**1**) or survive (**0**). "Respond with only the description followed by the single digit decision.

Heart rate data: row['heart rate']

Patient metadata: row['patient metadata']"

Experiment 2.6

"You are given NICU time-series data, patient metadata and a brief clinical description of the time series. Predict whether the infant will die in 7 days,or whether the infant will survive. Respond with **only** a single digit: '1' if the infant will die in 7 days, or '0' if the infant will survive.

Heart rate data: row['heart rate']

Patient metadata: row['patient metadata']

Clinical description: row['clinical description']"

5.4 Additional Results

Models	Recognition			Differentiation			OaR
	TS	TS + Metadata	Rank	TS	TS + Metadata	Rank	
GPT-4o	0.669	0.631	2.5	0.865	0.872	2	2.25
GPT-4o-V	0.671	0.639	1	0.887	0.904	1	1
Qwen-14B	0.623	0.619	4.5	0.743	0.788	4.5	4.5
ChatTS-14B	0.638	0.638	3	0.786	0.793	3.5	3.25
Qwen-7B	0.647	0.607	5.5	0.694	0.712	7	6.25
Qwen-7B-V	0.614	0.609	6.5	0.797	0.756	4	5.25
Phi-mini	0.540	0.612	7	0.678	0.708	8	7.5
Phi-mini-V	0.551	0.613	6	0.703	0.714	6	6

Table 3: Recognition & Differentiation Accuracy with with Sentence-BERTw/Cosine Similarity-based distractors. OaR = Overall rank

These results confirm that performance trends are robust to the choice of negative-sampling strategy. VLMs again lead across both Recognition and Differentiation, GPT-4o-Vision ranking highest overall. ChatTS-14B continues to outperform its base LLM (Qwen-14B), underscoring the benefit of temporal specialization, while smaller open models show limited gains from metadata.

5.5 Implementation Details

Experiments are run through the OpenAI GPT-4o API and model inference endpoints for Qwen and Phi and ChatTS models. Batching is used where possible to minimize API overhead. Inference is parallelized across NVIDIA A6000 GPUs on UVA's Rivanna HPC cluster for models requiring local deployment. Each experiment is repeated with both distractor sampling methods to ensure robustness of results.