1st ICLR Workshop on Time Series in the Age of Large Models

Summary

This workshop will delve into aspects of time series prediction and analysis in the age of large models. This workshop builds upon our successful track record of fostering community engagement around large models for time series. Our inaugural NeurIPS 2024 workshop demonstrated strong community interest, attracting 99 submissions and over 500 participants (~1000 registered interest via Whova). Submissions spanned the full spectrum of the field—from building time series foundation models and leveraging pre-trained models from other modalities, to real-world applications and deployment experiences. The rich discussions at NeurIPS 2024 revealed both significant opportunities and fundamental limitations in current approaches, directly informing the research questions we aim to address in this iteration. Building on this momentum, we also organized the successful ICML 2025 Workshop on Foundation Models for Structured Data, which broadened our perspective by connecting time series researchers with the tabular data community.

Focus and Innovation: For ICLR 2026, we are strategically refocusing to dive deeper into outstanding research questions that emerged from our previous workshops - particularly around agents, interpretability, and context-informed predictions. This iteration features an evolved organizing team and fresh speaker lineup, reflecting the field's rapid development. The nascent nature of large time series models makes this workshop particularly timely for ICLR 2026, as the community continues to establish foundational principles and explore novel applications in this emerging domain.

Organizer Expertise: The organizers bring extensive research experience and proven leadership in the time series foundation models domain, with diverse backgrounds from industry and academia. Collectively, we have led advances on 3 key dimensions: foundational model development– creating some of the first time series foundation models including Lag-Llama, Chronos, Moment, Moirai, and TimesFM, advanced applications– establishing initial frameworks for reasoning and agents in time series through MLZero and TimeSeriesGym, and rigorous evaluation and benchmarking using tools such as Context-is-Key, GIFT-Eval, TimeSeriesExam, and fev-bench. Beyond research contributions, our team has demonstrated success in organizing impactful workshops at premier venues, including the NeurIPS 2024 workshop on Time Series in the Age of Large Models, AAAI'24 Spring Symposium on Clinical Foundation Models, ICAIF'24 Foundation Models for Time Series: Exploring New Frontiers, and ICML'25 Workshop on Foundation Models for Structured Data. This combination of deep technical expertise and proven workshop leadership positions us to facilitate meaningful discussions and foster collaboration in this rapidly evolving field. The full list and bio of organizers is given at the end of the workshop proposal.

We describe below the problems we would like to address and the sub-problems where we expect advances to be made through this workshop.

Topics of the workshop

Context-informed foundation models for time series analysis: Over the past two years, several time series foundation models have been proposed in the literature [3-9, 35-37]. However, most prior works have focused on building foundation models for *univariate* time series forecasting. Although addressing univariate forecasting is important, most practical scenarios require additional context for enabling accurate predictions. For example, scenarios requiring *multivariate forecasting* and *forecasting with exogenous variables* (which may be static or dynamic) are ubiquitous in practice. With the advent of better natural language models, another possibility of *multimodal forecasting* has emerged [41-43], i.e., when the exogenous information comes from another modality such as text or images. We seek to understand the progress in these areas and the potential challenges in developing such foundation models which enable zero-shot inference with external context. We particularly encourage contributions of context-informed models that go beyond forecasting, tackling challenges in *anomaly detection* and *classification* that have applications in the industry (e.g., system failure and diagnosis) and healthcare [32].

Reasoning and time series ML agents: LLM-powered agents are transforming software engineering, machine learning, and data science by automating complex workflows [26-30]. In time series analysis, these agents have the potential to democratize ML by enabling domain experts to seamlessly integrate foundation models with classical methods. However, research on LLM agents for time series remains nascent [25, 31]. We welcome submissions on novel agents and scaffolds for time series modeling, as well as comprehensive benchmarks for evaluating agent performance [25, 31]. We would like to better understand which problems are better served by agents with time series models as tools versus developing tailored foundation models. Moreover, reasoning capabilities in time series models, both explicitly through approaches such as ReAct [33] and CodeAct [34] and implicit reasoning within foundation models represent a critical yet underexplored research direction [24, 29].

Benchmarks, datasets and tools for large time series models: A major bottleneck in advancing large time series models lies in the lack of standardized benchmarks, diverse large-scale datasets, and accessible tools for training and evaluation. This workshop welcomes contributions that propose unified benchmarking frameworks for evaluating large time series models. We are particularly interested in datasets that reflect real-world complexity, such as multivariate and context-informed time series tasks including multimodal time series datasets [15, 16]. We also encourage the creation of "live" benchmarks with data from sources that continuously update, which can be used to evaluate large time series models and LLM-based models without the risk of memorization [38]. Additionally, we highlight the importance of synthetic data generation for training large time series models [4], as the amount of high-quality public time series data is limited. An open challenge is how to effectively combine real and synthetic data during pretraining [23], where principled approaches to designing the pretraining mix remain largely unexplored. Furthermore, we encourage the release of open-source tools and libraries that facilitate large-scale pretraining, scalable data preprocessing, and efficient inference for time series models.

Interpretability of large time series models: As large time series models become increasingly prevalent, understanding their inner workings [17] and providing human-interpretable outcomes [18] are becoming both crucial and challenging. In this workshop, we aim to address the growing need for interpretability in large time series models. While these models demonstrate strong empirical performance across diverse tasks, their opaque nature limits trust, debugging, and adoption in high-stakes applications such as healthcare, finance, and climate modeling. We seek to explore methods that provide transparency into how these models make predictions, attribute importance to different time steps or features, or detect spurious correlations. This includes, but is not limited to, attention analysis [19], feature attribution techniques [20], symbolic distillation [21], and causal interpretability [22]. Additionally, we welcome contributions that examine how interpretability methods scale with model size and data heterogeneity, and how they differ across architectures (e.g., transformers vs. CNNs) and pretraining paradigms. Finally, we aim to bring the community together to discuss what interpretability should mean in the context of time series modeling, and to propose standardized evaluation setups and metrics for interpretability in this domain.

Evaluation and real-world applications of large time series models: As time series models grow in capability and move toward real-world impact across business verticals, evaluating progress becomes critical to ensure it translates to deployment. Recent work has exposed shortcomings in current metrics used for evaluation in time series prediction models [10], specifically in probabilistic prediction models [1, 2]. We seek contributions that deepen our understanding of the failure modes of these metrics and contributions developing new metrics for probabilistic prediction evaluation. We further invite contributions that delve deeper into specific real-world applications of time series predictions (such as retail or transportation): in understanding objectives important for each of these domains, to build better evaluation setups for time series models. Finally, we invite contributions that showcase the potential of large time series models in domains such as dynamical systems [39, 40], financial forecasting [12], human mobility forecasting [11], weather forecasting [13], clinical waveform interpretation [32] etc. Finally, we also encourage work that bridges the gap between forecasting and downstream decision-making, such as methods that integrate probabilistic predictions directly into optimization pipelines [14], with evaluations that reflect the end-to-end value of forecasts.

Invited Speakers

A list of invited speakers: All the following speakers have *agreed* to deliver an invited talk at the workshop.

Emily Fox is a Professor in the Department of Statistics and Department of Computer Science at Stanford University. Her research interests are in modeling complex time series arising in health, particularly from health wearables and

neuroimaging modalities. Prior to Stanford, she was the Amazon Professor of Machine Learning in the Paul G. Allen School of Computer Science & Engineering and Department of Statistics at the University of Washington. From 2018-2021, Emily led the Health AI team at Apple, where she was a Distinguished Engineer. Emily received an S.B. and Ph.D. from the Department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology (MIT). She has been awarded a Presidential Early Career Award for Scientists and Engineers (PECASE), Sloan Research Fellowship, ONR Young Investigator award, NSF CAREER award, and Leonard J. Savage Thesis Award in Applied Methodology.

Tim Januchowski is a Director of Engineering at Databricks and the Site Lead for the Databricks presence in Berlin. He oversees work on Unity Catalog and Object Storage. Before joining Databricks, he was the Director of Pricing Platform at Zalando SE, where he led the organization responsible for optimizing discounts for the Zalando wholesale business. This involves forecasting of demand heavily. Prior to Zalando, Tim led the time series science organization for Amazon Web Services' AI division. His teams built multiple AI services for AWS such as SageMaker, Forecast, Lookout for Metrics, and DevOps Guru, top-tier scientific publications, patents, and open source. Tim is a director at the International Institute of Forecasters, serves as a reviewer for the major ML venues, lectures at TU Munich & HPI, advises start-ups such as WhyLabs and Nixtla and runs the first and most popular forecasting course at Maven.

Caiming Xiong is the SVP of Salesforce AI research, where he leads efforts on foundation models, multimodal learning, and agent-based systems. His recent work focuses on building general-purpose agents and scalable AI frameworks for enterprise applications. He has developed influential models such as Moirai, CodeGen, BLIP, with publications in top venues including NeurIPS, ICML, ICLR, and ACL. His work has received multiple Best Paper Awards and over 58000 citations and an h-index of 104. At the workshop, he will talk about recent exploration in time series foundation models and their enterprise applications.

Marinka Zitnik is an Associate Professor at Harvard in the Department of Biomedical Informatics, an Associate Faculty at the Kempner Institute for the Study of Natural and Artificial Intelligence, Broad Institute of MIT and Harvard, and Harvard Data Science. Dr. Zitnik investigates foundations of AI to enhance scientific discovery and facilitate individualized diagnosis and treatment in medicine. Before joining Harvard, she was a postdoctoral scholar in Computer Science at Stanford University. She was also a member of the Chan Zuckerberg Biohub at Stanford. This research received best paper and research awards from International Society for Computational Biology, Bayer Early Excellence in Science Award, Amazon Faculty Research Award, Google Faculty Research Scholar Award, Roche Alliance with Distinguished Scientists Award, Sanofi iDEA-iTECH Award, Rising Star Award in Electrical Engineering and Computer Science (EECS), and Next Generation Recognition in Biomedicine. Dr. Zitnik co-founded Therapeutics Data Commons and is the faculty lead of the AI4Science initiative. Dr. Zitnik is the recipient of the 2022 Young Mentor Award at Harvard Medical School.

Yuyang (Bernie) Wang is a Principal Machine Learning Scientist in Amazon AI Labs, working mainly on large-scale probabilistic machine learning with its application in Forecasting. He received his PhD in Computer Science from Tufts University, MA, US and he holds an MS from the Department of Computer Science at Tsinghua University, Beijing, China. His research interests span statistical machine learning, numerical linear algebra, and random matrix theory. In forecasting, Yuyang has worked on all aspects ranging from practical applications to theoretical foundations.

Rose Yu is an Associate Professor at the University of California, San Diego, and a leading researcher at the intersection of large-scale spatiotemporal modeling and AI for science. Her research focuses on building scalable, principled deep learning methods for spatiotemporal data, particularly in integrating physics-inspired priors and data-driven techniques to enhance generalization and interpretability, with broad applications across climate science, transportation, and healthcare. Her work has been featured in top-tier venues including NeurIPS, ICML, and ICLR, and she has received numerous awards such as Presidential Early Career Award for Scientists and Engineers (PECASE), DARPA Young Faculty Award, ECASE Award, NSF CAREER Award, Hellman Fellowship and several Best Paper Awards. She was named as MIT Technology Review Innovators Under 35 in AI. At the workshop, her talk will be on "Token meets time: bridging LLMs with time series data".

Efforts made to ensure diversity

Efforts made to ensure diversity of organizers: The organizing team reflects diverse perspectives across multiple dimensions. Our team of 8 organizers includes 5 male and 3 organizers, from across 6 countries: Canada, China, Germany, India, Argentina, and the United States. The organizers represent a range of 3 academic institutions including

Carnegie Mellon University, the University of Montreal and University College London, as well as 5 companies such as Amazon, JP Morgan, Salesforce, ServiceNow, and Google, with significant presence in time series research. This diversity extends to career stages and expertise, with team members ranging from PhD students to senior research and applied scientists across different industry sectors. This varied composition ensures our event benefits from multiple viewpoints and experiences in planning and execution.

Efforts made to ensure demographic diversity of speakers: We have carefully curated a speaker lineup that ensures diverse representation across multiple dimensions. The lineup of 6 six speakers includes 3 male and 3 female leaders in time series research and applications, with cultural experiences from North America, Asia, and Europe. Our speakers balance academic and industry expertise, featuring three distinguished professors from top-tier universities (Stanford, Harvard, and UC San Diego) alongside time series researchers from major technology companies (Salesforce, AWS, and Databricks). They span different career stages and research focus areas, from distinguished professors to SVPs, covering AI for science, health applications, and foundation models. This diverse composition ensures our audience benefits from varied perspectives spanning theoretical foundations to enterprise applications across healthcare, climate science, and AI systems.

Efforts made to ensure diverse participants: To ensure a diverse range of participants, we will encourage submissions from underrepresented groups in AI and in interdisciplinary areas involving time series research. We will actively promote the workshop through relevant mailing lists, social media platforms (e.g., Twitter/X, LinkedIn, and specialized research communities), and direct outreach to labs and industry teams working on time series, machine learning, and interdisciplinary domains such as healthcare, finance, and climate science. Additionally, all workshop materials, including accepted papers, presentation videos, and panel discussions, will be made publicly available on the workshop website to ensure full virtual access and long-term impact.

Timeline and Agenda

Timeline of the workshop: All deadlines are in AoE.

Submission Deadline	February 7
Review Period	February 10 – 24
Paper Decisions	March 1, 2026

Agenda: Based on the previous edition of the workshop, we estimate the number of attendees in our workshop to be 500, and expect the number of submissions to be around 100, with an acceptance rate of 70-75%. The agenda below shows the planned schedule of the 6 keynotes (invited talks) and 4 contributed talks (oral papers), and poster sessions that facilitate discussion. Beyond the talks and poster sessions, one of the goals of the workshop is to build a community of researchers working on time series topics. We have planned a networking session that will facilitate this. The proposed program has sufficient time to (~ 4 hours) for semi-structured discussions.

09:00 – 09:10	Opening Remarks	13:35 – 13:55	Contributed Talks (x2)
09:10 – 09:45	Invited Talk 1 and Q&A	13:55 – 14:30	Invited Talk 4 and Q&A
09:45 – 10:05	Contributed Talks (x2)	14:30 – 15:30	Poster Session 2
10:05 – 10:40	Invited Talk 2 and Q&A	15:30 – 16:05	Invited Talk 5 and Q&A
10:40 – 12:00	Poster Session 1	16:05 – 16:35	Coffee & Networking Session
12:00 – 13:00	Lunch & Networking Session	16:35 – 17:10	Invited Talk 6 and Q&A
13:00 – 13:35	Invited Talk 3 and Q&A	17:10 – 17:30	Closing Remarks & Community Feedback

Dissemination: We will publish all accepted papers on our website, record and publish all the invited talks as well as the oral and spotlight presentations.

Submission Tracks

Submission Tracks: Submission Tracks: To foster broad participation and capture the rapidly evolving landscape of time series foundation models, our workshop welcomes submissions across two complementary tracks designed to bridge academic research and practical innovation.

(1) **Research Track (maximum 4 pages)** We invite research papers presenting novel theoretical insights, methodological advances, or empirical findings related to time series foundation models. To maintain focus on emerging work, papers already accepted at other venues are discouraged from submission.

(2) **Industry & Applications Track (maximum 2 pages)** Recognizing that significant innovation occurs beyond traditional academic settings, we welcome submissions from industry practitioners, open-source contributors, and applied researchers. Aligned with ICLR's participatory initiative for concise papers, this track prioritizes practical impact and real-world insights over strict novelty requirements. This track encompasses: (a) Implementation and evaluation of simple but unpublished ideas, (b) Self-contained theoretical results or proofs-of-concept, (c) Follow-up experiments or re-analyses of existing work, (d) Fresh perspectives, critiques, or novel interpretations of published research, (e) Real-world deployment experiences and lessons learned, (f) Open-source tools, libraries, datasets, and frameworks, (g) Large-scale applications and production systems.

Both tracks welcome work-in-progress submissions and are non-archival. We particularly encourage participation from student first authors and researchers outside the traditional ML conference circuit, fostering an inclusive environment for diverse perspectives on time series foundation models. Notably, we will welcome short or tiny papers in both tracks.

Organizers

1. Abdul Fatir Ansari, **Amazon Web Services**, abdulfatirs@gmail.com, google scholar

Abdul Fatir Ansari is a Senior Applied Scientist at Amazon Web Services, focusing on time series forecasting. He is the lead developer of the Chronos family of time series foundation models, which ranked among the most downloaded models on HuggingFace in 2024. He received his PhD in 2022 from the National University of Singapore under the supervision of Prof. Harold Soh. During his PhD, he worked on deep generative models for images and time series, receiving the Dean's Award for his research. He has published in and reviewed for top machine learning conferences including ICML, ICLR, and NeurIPS, and was twice recognized as a top reviewer. He co-organized the "Time Series in the Age of Large Models" and "Foundation Models for Structured Data" workshops at NeurIPS 2024 and ICML 2025, respectively. His research interests include time series analysis and generative modeling.

2. Arjun Ashok, **University of Montreal, Mila-Quebec AI Institute** and **ServiceNow Research**, arjun.ashok.psg@gmail.com, google scholar

Arjun Ashok is a PhD Student at the University of Montreal and Mila-Quebec AI Institute, advised by Irina Rish and Alexandre Drouin. He is also a Visiting Researcher at ServiceNow Research, Montreal, Canada. He is the co-creator of the Lag-Llama foundation model for forecasting and the Context-is-Key benchmark for context-aided forecasting, among other works. His research interests lie in time series forecasting and decision-making, with a focus on designing scalable general-purpose models for time series prediction tasks. He has given multiple talks on his work on flexible time series prediction models and time series foundation models. He was a co-organizer of the "Time Series in the Age of Large Models" workshop at NeurIPS 2024. He serves as a regular reviewer for machine learning and forecasting conferences and has been recognized as a top reviewer.

3. Elizabeth Fons, **JP Morgan AI Research and University College London**, elizabeth.fons@jpmorgan.com, google scholar

Elizabeth is a Research Lead in the AI Research Group at JP Morgan, specializing in modeling and understanding time series data, with an emphasis on generative methods, imputation, and leveraging large language models for time series interpretation and reporting. She is also a Lecturer at University College London. Elizabeth received her PhD in

Computer Science from University of Manchester, where her research, conducted in collaboration with AllianceBernstein, focused on machine learning applications for time series analysis in finance. Prior to her PhD, Elizabeth obtained an MSc in Physics from the University of Buenos Aires, Argentina.

4. Xiyuan Zhang, **AWS AI,** xiyuanz@amazon.com, google scholar

Xiyuan is an Applied Scientist at Amazon Web Services working on time series forecasting, especially on pre-training and multimodal analysis. She is the co-author of Chronos, the most downloaded time series foundation model on HuggingFace. Xiyuan earned her PhD in Computer Science from the University of California, San Diego, advised by Prof. Rajesh K. Gupta and Prof. Jingbo Shang. She has published and received Top Reviewer Award in top machine learning conferences such as NeurIPS, and has co-organized the "Foundation Models for Structured Data" workshop at ICML 2025. Xiyuan is a recipient of the Qualcomm Innovation Fellowship and has been recognized as a Cyber-Physical-System (CPS) Rising Star. Her broad research interests lie in machine learning for sequential data (time series, spatiotemporal, tabular, NLP), with practical impact including healthcare, IoT, climate science and beyond.

5. Chenghao Liu, **Salesforce AI Research**, chenghao.liu@salesforce.com, google scholar

Chenghao is a Lead Applied Scientist in Salesforce AI Research and leads the time series research team to support IT Operations. He received B.S. and PH.D. degrees in Computer Science and Technology from Zhejiang University. His research interests include machine learning, data mining and their applications. He has published over 90 top-ranked AI conference and journal papers, had multiple Oral/Spotlight Papers at NeurIPS, ICML, KDD, ICLR, AAAI and IJCAI, etc. Currently, his research focuses on time series foundation model, time series representation learning and causal analysis.

6. Mononito Goswami, **AWS Agentic AI**, mononito@amazon.com, google scholar

Mononito is an Applied Scientist in the Agentic AI organization in Amazon Web Services, focusing on building pragmatic agents which can model structured data. He recently earned his Ph.D. from Carnegie Mellon University, where his research focused on developing foundation models for structured data. He has led the development of widely used foundation models, including MOMENT, a foundation model for time series that has garnered over 2 million downloads on HuggingFace and attracted more than $2 million in research funding. In 2021, he was awarded the Center for Machine Learning and Health (CMLH) fellowship. His research has been published in premier machine learning conferences including NeurIPS, ICLR, and ICML. He serves as a regular reviewer for these venues and has been recognized as a top reviewer. Mononito co-organized the AAAI 2024 Spring Symposium on Clinical Foundation Models and the Foundation Models for Structured Data workshop at ICML 2025. His research interests lie at the intersection of LLM agents, foundation models and structured data, with a particular focus on developing practical machine learning solutions for healthcare and education applications.

7. Xinyu Li, **Carnegie Mellon University**, xinyul2@andrew.cmu.edu, google scholar

Xinyu is a PhD student at Carnegie Mellon University, advised by Prof. Artur Dubrawski. Her research centers on LLM-based agents, with a focus on developing scalable benchmarks and general agentic frameworks for time series machine learning tasks. She has also gained industry research experience at Microsoft and Amazon, and her work has been presented at multiple ML conferences and workshops. Xinyu also co-organized the AAAI 2024 Spring Symposium on Clinical Foundation Models.

8. Yichen Zhou, **Google Research**, yichenzhou@google.com, Google Scholar

Yichen Zhou is a Research Engineer at Google Research. He received his PhD in Stats and MS in CS from Cornell University, and BS in Mathematics from Tsinghua University. His current research focuses on foundation models for structured data. He is the co-author of the TimesFM family of time series foundation models. He has published in and reviewed for top machine learning and statistics venues including ICLR, NeurIPS, ICML, and JMLR.

Note: No other workshops will be concurrently organized by any of the organizers.

Reviewing Process

Program Committee and Reviewing Process: We will use OpenReview to manage submissions and recruit reviewers through an open application form distributed with the call for papers. Our selection criteria require reviewers to have at least one published work on time series prediction in a top-tier ML conference or journal within the past three years and be actively working in the field. In previous iterations of this workshop and other workshops, we successfully recruited over 90 qualified reviewers. This enabled us to provide three reviews per paper on average while assigning only three papers per reviewer. We will build on this established network and process to ensure thorough, high-quality reviews for all submissions.

How conflict of interest will be managed between the organizers: To ensure fair and unbiased evaluation, organizers will recuse themselves from reviewing submissions under the following circumstances: (1) Institutional conflicts: Organizers will not evaluate submissions from authors with whom they share current or recent institutional affiliations (within the past three years), including educational and professional connections. (2) Collaborative relationships: Organizers will abstain from assessing work by authors with whom they have had research collaborations in the last three years, including co-authored publications, jointly organized events, or other professional partnerships. (3) Personal relationships: Organizers will not participate in the review process for submissions from authors with whom they have close personal ties.

References

[1] Koochali, A., Schichtel, P., Dengel, A., & Ahmed, S. (2022). Random noise vs. state-of-the-art probabilistic forecasting methods: A case study on CRPS-Sum discrimination ability. Applied Sciences, 12(10), 5104.

[2] Marcotte, É., Zantedeschi, V., Drouin, A., & Chapados, N. (2023, July). Regions of reliability in the evaluation of multivariate probabilistic forecasts. In International Conference on Machine Learning (pp. 23958-24004). PMLR.

[3] Rasul, K., Ashok, A., Williams, A. R., Khorasani, A., Adamopoulos, G., Bhagwatkar, R., … & Rish, I. (2023). Lag-LLama: Towards foundation models for time series forecasting. arXiv preprint arXiv:2310.08278.

[4] Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., … & Wang, Y. (2024). Chronos: Learning the language of time series. arXiv preprint arXiv:2403.07815.

[5] Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., & Sahoo, D. (2024). Unified training of universal time series forecasting transformers. arXiv preprint arXiv:2402.02592.

[6] Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., & Dubrawski, A. (2024). MOMENT: A Family of Open Time-series Foundation Models. arXiv preprint arXiv:2402.03885.

[7] Das, A., Kong, W., Sen, R., & Zhou, Y. (2023). A decoder-only foundation model for time-series forecasting. arXiv preprint arXiv:2310.10688.

[8] Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., … & Wen, Q. (2023, October). Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. In The Twelfth International Conference on Learning Representations.

[9] Cao, D., Jia, F., Arik, S. O., Pfister, T., Zheng, Y., Ye, W., & Liu, Y. (2023, October). TEMPO: Prompt-based Generative Pre-trained Transformer for Time Series Forecasting. In The Twelfth International Conference on Learning Representations.

[10] Hewamalage, H., Ackermann, K., & Bergmeir, C. (2023). Forecast evaluation for data scientists: common pitfalls and best practices. Data Mining and Knowledge Discovery, 37(2), 788-832.

[11] Xue, H., Voutharoja, B. P., & Salim, F. D. (2022, November). Leveraging language foundation models for human mobility forecasting. In Proceedings of the 30th International Conference on Advances in Geographic Information Systems (pp. 1-9).

[12] Yu, X., Chen, Z., Ling, Y., Dong, S., Liu, Z., & Lu, Y. (2023). Temporal Data Meets LLM—Explainable Financial Time Series Forecasting. arXiv preprint arXiv:2306.11025.

[13] Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., & Grover, A. (2023). ClimaX: A foundation model for weather and climate. arXiv preprint arXiv:2301.10343.

[14] Makridakis, S., Spiliotis, E., Hollyman, R., Petropoulos, F., Swanson, N., & Gaba, A. (2024). The M6 forecasting competition: Bridging the gap between forecasting and investment decisions. International Journal of Forecasting.

[15] Liu, H., Xu, S., Zhao, Z., Kong, L., Kamarthi, H., Sasanur, A. B., … & Prakash, B. A. (2024). Time-MMD: A new multi-domain multimodal dataset for time series analysis. arXiv preprint arXiv:2406.08627.

[16] Zhou, X., Wang, W., Baldán, F. J., Buntine, W., & Bergmeir, C. (2025). MoTime: A Dataset Suite for Multimodal Time Series Forecasting. arXiv preprint arXiv:2505.15072.

[17] Wiliński, M., Goswami, M., Żukowska, N., Potosnak, W., & Dubrawski, A. (2024). Exploring representations and interventions in time series foundation models. arXiv preprint arXiv:2409.12915.

[18] Ismail, A. A., Gunady, M., Corrada-Bravo, H., & Feizi, S. (2020). Benchmarking deep learning interpretability in time series predictions. Advances in Neural Information Processing Systems, 33, 6441-6452.

[19] Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT look at? An analysis of BERT's attention. arXiv preprint arXiv:1906.04341.

[20] Zhou, Y., Booth, S., Ribeiro, M. T., & Shah, J. (2022, June). Do feature attribution methods correctly attribute features? In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 9, pp. 9623-9633).

[21] Acharya, K., Velasquez, A., & Song, H. H. (2024). A survey on symbolic knowledge distillation of large language models. IEEE Transactions on Artificial Intelligence.

[22] Zhao, Q., & Hastie, T. (2021). Causal interpretations of black-box models. Journal of Business & Economic Statistics, 39(1), 272-281.

[23] Liu, X., Aksu, T., Liu, J., Wen, Q., Liang, Y., Xiong, C., Savarese, S., Sahoo, D., Li, J., & Liu, C. (2025). Empowering time series analysis with synthetic data: A survey and outlook in the era of foundation models. arXiv preprint arXiv:2503.11411.

[24] Cai, Y., Choudhry, A., Goswami, M., & Dubrawski, A. (2024). TimeSeriesExam: A time series understanding exam. NeurIPS 2024 Workshop on Time Series in the Age of Large Models.

[25] Cai, Y., Li, X., Goswami, M., Wiliński, M., Welter, G., & Dubrawski, A. (2025). TimeSeriesGym: A scalable benchmark for (time series) machine learning engineering agents. arXiv preprint arXiv:2505.13291.

[26] Chan, J. S., Chowdhury, N., Jaffe, O., Aung, J., Sherburn, D., Mays, E., … & Weng, L. (2024). MLE-Bench: Evaluating machine learning agents on machine learning engineering. arXiv preprint arXiv:2410.07095.

[27] Jiang, Z., Schmidt, D., Srikanth, D., Xu, D., Kaplan, I., Jacenko, D., & Wu, Y. (2025). AIDE: AI-driven exploration in the space of code. arXiv preprint arXiv:2502.13138.

[28] Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. (2023). SWE-Bench: Can language models resolve real-world GitHub issues? arXiv preprint arXiv:2310.06770.

[29] Potosnak, W., Challu, C., Goswami, M., Olivares, K. G., Wiliński, M., Żukowska, N., & Dubrawski, A. (2025). Investigating compositional reasoning in time series foundation models. arXiv preprint arXiv:2502.06037.

[30] Wang, X., Li, B., Song, Y., Xu, F. F., Tang, X., Zhuge, M., … & Neubig, G. (2024, October). OpenHands: An open platform for AI software developers as generalist agents. In The Thirteenth International Conference on Learning Representations.

[31] Ye, W., Zhang, Y., Yang, W., Tang, L., Cao, D., Cai, J., & Liu, Y. (2024). Beyond forecasting: Compositional time series reasoning for end-to-end task execution. arXiv preprint arXiv:2410.04047.

[32] Cai, Y., Goswami, M., Choudhry, A., Srinivasan, A., & Dubrawski, A. (2023). JOLT: Jointly learned representations of language and time-series. In Deep Generative Models for Health Workshop, NeurIPS 2023.

[33] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., & Cao, Y. ReAct: Synergizing reasoning and acting in language models. In The Eleventh International Conference on Learning Representations.

[34] Wang, X., Chen, Y., Yuan, L., Zhang, Y., Li, Y., Peng, H., & Ji, H. (2024, July). Executable code actions elicit better LLM agents. In Forty-first International Conference on Machine Learning.

[35] Shi, X., Wang, S., Nie, Y., Li, D., Ye, Z., Wen, Q., & Jin, M. (2025). Time-MOE: Billion-scale time series foundation models with mixture of experts. ICLR 2025.

[36] Ekambaram, V., Jati, A., Dayama, P., Mukherjee, S., Nguyen, N., Gifford, W. M., … & Kalagnanam, J. (2024). Tiny Time Mixers (TTM): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. Advances in Neural Information Processing Systems, 37, 74147-74181.

[37] Cohen, B., Khwaja, E., Doubli, Y., Lemaachi, S., Lettieri, C., Masson, C., … & Abou-Amal, O. (2025). This Time is Different: An observability perspective on time series foundation models. arXiv preprint arXiv:2505.14766.

[38] White, C., Dooley, S., Roberts, M., Pal, A., Feuer, B., Jain, S., … & Goldblum, M. LiveBench: A challenging, contamination-free LLM benchmark. In The Thirteenth International Conference on Learning Representations.

[39] Zhang, Y., & Gilpin, W. (2024). Zero-shot forecasting of chaotic systems. arXiv preprint arXiv:2409.15771.

[40] Lai, J., Bao, A., & Gilpin, W. (2025). PANDA: A pretrained forecast model for universal representation of chaotic

dynamics. arXiv preprint arXiv:2505.13755.

[41] Williams, A. R., Ashok, A., Marcotte, É., Zantedeschi, V., Subramanian, J., Riachi, R., ... & Drouin, A. (2024). Context is key: A benchmark for forecasting with essential textual information. arXiv preprint arXiv:2410.18959.

[42] Liu, H., Kamarthi, H., Zhao, Z., Xu, S., Wang, S., Wen, Q., ... & Prakash, B. A. (2025). How can time series analysis benefit from multiple modalities? a survey and outlook. arXiv preprint arXiv:2503.11835.

[42] Jiang, Y., Ning, K., Pan, Z., Shen, X., Ni, J., Yu, W., ... & Song, D. (2025). Multi-modal Time Series Analysis: A Tutorial and Survey. arXiv preprint arXiv:2503.13709.