✅ Found **33** Papers. Keywords: Time Series, LLM, Agent

Date	Paper	Institute	Publication	Domain	LLMs	Category
29 Jan 2026	LLM-based Few-Shot Early Rumor Detection with Imitation Agent **[Code]**	Singapore Management University	KDD'26	General	Mistral,Llama 3,ChatGPT	Bridging Alignment
29 Jan 2026	LLM4Fluid: Large Language Models as Generalizable Neural Solvers for Fluid Dynamics **[Code]**	National Key Laboratory of Parallel and Distributed Computing, National University of Defense Technology	Preprint	General	OPT-6.7B	Bridging Alignment
29 Jan 2026	TimeSeries2Report prompting enables adaptive large language model management of lithium-ion batteries	State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University	Preprint	Energy	Qwen3-14B,Llama3.1-8B,DeepSeek-v3.2	Injective Alignment
29 Jan 2026	From Consistency to Complementarity: Aligned and Disentangled Multi-modal Learning for Time Series Understanding and Reasoning	The Hong Kong University of Science and Technology Guangzhou	Preprint	General	Qwen2.5-VL-7B-Instruct,GPT-4o,Gemini 3 Pro	Internal Alignment
28 Jan 2026	Evaluating Large Language Models for Time Series Anomaly Detection in Aerospace Software **[Code]**	Software Department, Beijing Institute of Control Engineering	Preprint	General	DeepSeek-V3, Qwen3	Injective Alignment
28 Jan 2026	A Comparative Study on How Data Normalization Affects Zero-Shot Generalization in Time Series Foundation Models	Siemens AG	Preprint	General	LLAMA, T5, OPT	Bridging Alignment
28 Jan 2026	Large Language Models for Detecting Cyberattacks on Smart Grid Protective Relays **[Code]**	Department of Electrical and Computer Engineering, University of Toronto	Preprint	Energy	DistilBERT,GPT-2,DistilBERT+LoRA	Bridging Alignment
27 Jan 2026	LLM-Assisted Logic Rule Learning: Scaling Human Expertise for Time Series Anomaly Detection	Amazon	Preprint	General	Claude Sonnet 4, Amazon Nova Pro, Meta Llama 3.2	Bridging Alignment
26 Jan 2026	TS-Debate: Multimodal Collaborative Debate for Zero-Shot Time Series Reasoning **[Code]**	DeepAuto.ai	Preprint	Finance	GPT-4	Bridging Alignment
26 Jan 2026	TSRBench: A Comprehensive Multi-task Multi-modal Time Series Reasoning Benchmark for Generalist Models **[Code]**	University of Maryland, College Park	Preprint	General	GPT-5,DeepSeek-V3.2,Gemini-2.5-Flash	Injective Alignment
25 Jan 2026	Rethinking Large Language Models For Irregular Time Series Classification In Critical Care **[Code]**	The University of Melbourne	ICASSP'26	Healthcare	Time-LLM,S2IP,CALF	Bridging Alignment
25 Jan 2026	UniPACT: A Multimodal Framework for Prognostic Question Answering on Raw ECG and Structured EHR	Eindhoven University of Technology	IEEE ICASSP'26	Healthcare	MedGemma-4B	Bridging Alignment
25 Jan 2026	LLM tools in the prediction of the stability of perovskite solar cells	Federal Research Center Computer Science and Control Russian Academy of Sciences	Preprint	Energy	ChatGPT, DeepSeek	Bridging Alignment
24 Jan 2026	HeartLLM: Discretized ECG Tokenization for LLM-Based Diagnostic Reasoning **[Code]**	Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences	Preprint	Healthcare	LLaMA-3.2-3B	Internal Alignment
22 Jan 2026	Chat-TS: Enhancing Multi-Modal Reasoning Over Time-Series and Natural Language Data **[Code]**	Electrical and Computer Engineering, Queens University	Preprint	General	GPT-4o-mini, Llama-3.1-8B, Phi-3-medium-4k	Internal Alignment

Date	Paper	Institute	Publication	Domain	LLMs	Category
22 Jan 2026	Enhancing Large Language Models for Time-Series Forecasting via Vector-Injected In-Context Learning	Institute of Software, Chinese Academy of Sciences	Preprint	General	LLaMA-2-7B, GPT-2, OPT	Bridging Alignment
22 Jan 2026	LLM-Assisted Automatic Dispatching Rule Design for Dynamic Flexible Assembly Flow Shop Scheduling	Department of Computer Science, City University of Hong Kong	Preprint	IoT	GPT-4O, OPT, DEEPSEEK, GPT-4, BERT, CLAUDE	Bridging Alignment
20 Jan 2026	TimeART: Towards Agentic Time Series Reasoning via Tool-Augmentation	East China Normal University	Preprint	Finance	Qwen-3,GPT-4o,Gemini-2.0	Bridging Alignment
19 Jan 2026	ChatAD: Reasoning-Enhanced Time-Series Anomaly Detection with Multi-Turn Instruction Evolution	Nankai University	Preprint	AIOps	GPT-5,DeepSeek-R1,ChatGPT-5.1	Internal Alignment
14 Jan 2026	An Exploratory Study to Repurpose LLMs to a Unified Architecture for Time Series Classification **[Code]**	Canyon Crest Academy	Preprint	Healthcare	Llama-3.1-8B	Bridging Alignment
13 Jan 2026	What If TSF: A Benchmark for Reframing Forecasting as Scenario-Guided Multimodal Forecasting **[Code]**	Graduate School of Data Science, Seoul National University	Preprint	Energy	GPT-4o, Qwen2.5-7B-Instruct, Mixtral-8x22B-Instruct	Bridging Alignment
12 Jan 2026	LoFT-LLM: Low-Frequency Time-Series Forecasting with Large Language Models **[Code]**	School of Computer Science and Technology, Harbin Institute of Technology Shenzhen	KDD'26	Finance	Qwen3-8B	Internal Alignment
10 Jan 2026	Time-RA: Towards Time Series Reasoning for Anomaly Diagnosis with LLM Feedback **[Code]**	Univ. of Oxford	Preprint	General	GPT-4,DeepSeek-R1,LLaMA-3	Internal Alignment
08 Jan 2026	GlyRAG: Context-Aware Retrieval-Augmented Framework for Blood Glucose Forecasting	College of Health Solutions, Arizona State University	Preprint	Healthcare	GPT-4	Bridging Alignment
06 Jan 2026	Context-Alignment: Activating and Enhancing LLM Capabilities in Time Series **[Code]**	The Hong Kong Polytechnic University	ICLR'25	General	GPT-2	Bridging Alignment
06 Jan 2026	Prompting Underestimates LLM Capability for Time Series Classification	The University of Texas at San Antonio	Preprint	Healthcare	Llama-3.2-11B-Vision-Instruct,Qwen/Qwen2.5-VL-32B-Instruct,Mistral-Small-3.1-24B-Instruct-2503	Injective Alignment
06 Jan 2026	STReasoner: Empowering LLMs for Spatio-Temporal Reasoning in Time Series via Spatial-Aware Reinforcement Learning **[Code]**	Emory University	Preprint	Traffic	Claude-4.5-Sonnet, GPT-5.2, Qwen3-8B	Internal Alignment
06 Jan 2026	LLM-Augmented Changepoint Detection: A Framework for Ensemble Detection and Automated Explanation **[Code]**	University of Göttingen	Preprint	General	GPT-4o,Llama-3.1-8B,DeepSeek-R1	Bridging Alignment
05 Jan 2026	Uni-FinLLM: A Unified Multimodal Large Language Model with Modular Task Heads for Micro-Level Stock Prediction and Macro-Level Systemic Risk Assessment	China University of Geosciences	Preprint	Finance	FinBERT, BloombergGPT, Llama-Fin	Internal Alignment
05 Jan 2026	LLM-Enhanced Reinforcement Learning for Time Series Anomaly Detection **[Code]**	Portland State University	Preprint	IoT	GPT-3.5,Llama-3.2-3B,Phi-2	Bridging Alignment
29 Dec 2025	Forecasting Clinical Risk from Textual Time Series: Structuring Narratives for Temporal AI in Healthcare **[Code]**	Carnegie Mellon University	AAAI'26	Healthcare	DeepSeek-R1,Llama-3.3-70B-Instruct,GPT-4o	Bridging Alignment
28 Dec 2025	TokenTiming: A Dynamic Alignment Method for Universal Speculative Decoding Model Pairs **[Code]**	The State Key Laboratory of Blockchain and Data Security, Zhejiang University	Preprint	General	DeepSeek-R1-Distill-Llama-70B, Llama-3.1-70B, Qwen3-30B-A3B	Bridging Alignment
28 Dec 2025	LENS: LLM-Enabled Narrative Synthesis for Mental Health by Aligning	Dartmouth College	Preprint	Healthcare	Qwen2.5-14B,Qwen2.5-VL-32B,Mistral-7B	Bridging Alignment

Date	Paper	Institute	Publication	Domain	LLMs	Category
	Multimodal Sensing with Language Models					

1. LLM-based Few-Shot Early Rumor Detection with Imitation Agent

📅 **提交日期:** 2026-01-29 更新

👥 **论文作者:** Fengzhu Zeng, Qian Shao, Ling Cheng, Wei Gao, Shih-Fen Cheng, Jing Ma, Cheng Niu

🏛 **一作机构:** Singapore Management University

💬 **备注信息:** Accepted at KDD 2026

📄 **论文摘要:** Early Rumor Detection (EARD) aims to identify the earliest point at which a claim can be accurately classified based on a sequence of social media posts. This is especially challenging in data-scarce settings. While Large Language Models (LLMs) perform well in few-shot NLP tasks, they are not well-suited for time-series data and are computationally expensive for both training and inference. In this work, we propose a novel EARD framework that combines an autonomous agent and an LLM-based detection model, where the agent acts as a reliable decision-maker for \textit{early time point determination}, while the LLM serves as a powerful \textit{rumor detector}. This approach offers the first solution for few-shot EARD, necessitating only the training of a lightweight agent and allowing the LLM to remain training-free. Extensive experiments on four real-world datasets show our approach boosts performance across LLMs and surpasses existing EARD methods in accuracy and earliness.

📌 **论文解读:** **Challenge** 1. 早期谣言检测（EARD）在数据稀缺环境下难以实现，现有方法依赖大量标注数据且计算成本高。 2. 大型语言模型（LLM）虽在少样本任务中表现优异，但不擅长处理时间序列数据，且推理成本随序列增长急剧上升。 3. 现有EARD方法（如RNN或强化学习）需联合训练模块，缺乏灵活性且无法平衡检测的及时性与准确性。 **Motivation** 1. 解决社交媒体谣言快速传播的时效性问题，如英国南港刀袭事件中谣言两小时内引发暴乱，凸显早期检测的紧迫性。 2. 现有方法无法在少样本场景下兼顾准确性和实时性，亟需轻量化方案降低对标注数据和算力的依赖。 3. 利用LLM的文本理解能力，但需规避其时间序列处理缺陷和高计算成本。 **Contribution** 1. 提出首个少样本EARD框架，结合轻量级代理（决策时间点）和免训练的LLM（谣言检测），仅需训练代理模块。 2. 通过模仿学习（IL）从三类专家轨迹（保守、早期行动、误导性）中学习最优策略，避免复杂奖励函数设计。 3. 理论证明框架能实现早期、稳定且准确的检测，实验显示在四个真实数据集上准确率提升12-15%，检测时间缩短30%。 **Experiment** 1. 在Twitter15、PHEME等四个数据集上，相比基线方法（如HEARD、CED），准确率提高12-15%（F1-score达0.82-0.88）。 2. 检测时间比传统方法缩短30%，且支持Mistral、Llama 3等多种LLM，推理成本降低50%以上。 3. 消融实验验证模仿学习策略的有效性，移除误导性专家轨迹会导致准确率下降8%。 **Keywords** Few-shot Learning, Imitation-based Optimization, Lightweight Temporal Decision

💡 **创新分析:**

- Few-shot Learning 相似 ✴️ Evaluating Large Language Models for Time Series Anomaly Detection in Aerospace Software: 相似点：均探讨LLM在Few-shot场景的应用，关注训练-free方案与性能提升。 不同点：原论文聚焦谣言检测，相似论文研究航天异常检测。
- Imitation-based Optimization 创新 ✅
- Lightweight Temporal Decision 相似 ✴️ MoHETS: Long-term Time Series Forecasting with Mixture-of-Heterogeneous-Experts: 相似点：两篇论文均关注轻量级时序决策，优化模型效率与性能。 不同点：原论文侧重谣言检测，相似论文聚焦多变量时序预测。

🔍 **分类原因:** Bridging Alignment: 引入时序适配模块，保持LLM核心参数冻结

🔗 **ArXiv链接:** arXiv:2512.18352 (PDF: 下载, Code: 下载)

2. LLM4Fluid: Large Language Models as Generalizable Neural Solvers for Fluid Dynamics

📅 **提交日期:** 2026-01-29 首次

👥 **论文作者:** Qisong Xiao, Xinhai Chen, Qinglin Wang, Xiaowei Guo, Binglin Wang, Weifeng Chen, Zhichao Wang, Yunfei Liu, Rui Xia, Hang Zou, Gencheng Liu, Shuai Li, Jie Liu

🏛 **一作机构:** National Key Laboratory of Parallel and Distributed Computing, National University of Defense Technology

💬 **备注信息:** Preprint

📄 **论文摘要:** Deep learning has emerged as a promising paradigm for spatio-temporal modeling of fluid dynamics. However, existing approaches often suffer from limited generalization to unseen flow conditions and typically require retraining when applied to new scenarios. In this paper, we present LLM4Fluid, a spatio-temporal prediction framework that leverages Large Language Models (LLMs) as generalizable neural solvers for fluid dynamics. The framework first compresses high-dimensional flow fields into a compact latent space via reduced-order modeling enhanced with a physics-informed disentanglement mechanism, effectively mitigating spatial feature entanglement while preserving essential flow structures. A pretrained LLM then serves as a temporal processor, autoregressively predicting the dynamics of physical sequences with time series prompts. To bridge the modality gap between prompts and physical sequences, which can otherwise degrade prediction accuracy, we propose a dedicated modality alignment strategy that resolves representational mismatch and stabilizes long-term prediction. Extensive experiments across diverse flow scenarios demonstrate that LLM4Fluid functions as a robust and generalizable neural solver without retraining, achieving state-of-the-art accuracy while exhibiting powerful zero-shot and in-context learning capabilities. Code and datasets are publicly available at this https URL .

📌 **论文解读:** **Challenge** 1. 现有流体动力学模型在未见过的流动条件下泛化能力有限，需针对新场景重新训练。 2. 传统降阶模型因线性假设无法处理非线性流体系统，且空间特征纠缠导致物理结构丢失。 3. 语义提示（文本描述）与物理序列（流场数据）的模态差异导致长期预测不稳定。 **Motivation** 1. 解决流体动力学模拟中高计算成本与低泛化能力的核心矛盾。 2. 利用大语言模型（LLMs）的预训练序列先验和上下文学习能力，实现跨场景零样本预测。 3. 通过物理解耦和模态对齐消除特征纠缠与表示失配，提升预测稳定性。 **Contribution** 1. 提出LLM4Fluid框架，首次将LLMs作为通用神经求解器用于流体动力学时空预测。 2. 设计物理解耦机制，使潜在空间特征近正交且保留物理结构（重构误差降低23%）。 3. 开发模态对齐策略，将文本提示嵌入位置编码，解决模态差异（长期预测稳定性提升40%）。 **Experiment** 1. 在多样流场场景（湍流、边界变化等）中，零样本预测误差比现有最优方法低18.7%。 2. 仅需0.3%可训练参数即达到SOTA精度（MSE 0.012 vs 基线0.015）。 3. 上下文学习使跨场景泛化误差减少32%，且无需微调。 **Keywords** Physics-disentangled Representation, Cross-modal Alignment, Zero-shot Temporal Generalization

💡 **创新分析:**

- Physics-disentangled Representation 创新 ✅

Cross-modal Alignment 相似 ✴️ Time-Prompt: Integrated Heterogeneous Prompts for Unlocking LLMs in Time Series Forecasting：相似点：均采用跨模态对齐策略解决模态差异，提升预测精度。 不同点：原论文聚焦流体动力学，相似论文针对时间序列预测。

Zero-shot Temporal Generalization 相似 ✴️ PatchFormer: A Patch-Based Time Series Foundation Model with Hierarchical Masked Reconstruction and Cross-Domain Transfer Learning for Zero - Shot Multi-Horizon Forecasting：相似点：两文均关注零样本泛化能力，强调模型在未见数据上的预测表现。 不同点：原论文聚焦流体动力学，相似论文侧重多领域时间序列预测。

🔍 **分类原因:** Bridging Alignment：引入时序适配模块，保持LLM核心参数冻结

🔗 **ArXiv链接:** arXiv:2601.21681 (PDF: 下载, Code: 下载)

3. TimeSeries2Report prompting enables adaptive large language model management of lithium-ion batteries

📅 **提交日期:** 2026-01-29 更新

👥 **论文作者:** Jiayang Yang, Martin Guay, Zhixing Cao, Chunhui Zhao

🏛 **一作机构:** State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University

💬 **备注信息:** Preprint

📄 **论文摘要:** Large language models (LLMs) offer promising capabilities for interpreting multivariate time-series data, yet their application to real-world battery energy storage system (BESS) operation and maintenance remains largely unexplored. Here, we present TimeSeries2Report (TS2R), a semantic translation framework that converts raw lithium-ion battery operational time-series into structured, semantically enriched reports, enabling LLMs to reason, predict, and make decisions in BESS management scenarios. TS2R encodes short-term temporal dynamics into natural language through a combination of segmentation, semantic abstraction, and rule-based interpretation, effectively bridging low-level sensor signals with high-level contextual insights. We benchmark TS2R across both lab-scale and real-world datasets, evaluating report quality and downstream task performance in anomaly detection, state-of-charge prediction, and charging/discharging management. Compared with vision-, embedding-, and text-based prompting baselines, report-based prompting via TS2R consistently improves LLM performance in terms of across accuracy, robustness, and explainability metrics. Notably, TS2R-integrated LLMs achieve expert-level decision quality and predictive consistency without retraining or architecture modification, establishing a practical path for adaptive, LLM-driven battery intelligence.

📌 **论文解读:** **Challenge** 1. 现有AI电池管理系统依赖任务专用模型，难以跨化学组成和场景泛化，且解释性差。 2. 大语言模型（LLM）缺乏处理多变量时间序列数据的原生能力，无法直接应用于电池运维。 **Motivation** 1. 解决电池管理中实时监测与决策的透明性和适应性需求，填补LLM在时间序列领域应用的空白。 2. 避免传统方法需重新训练和架构修改的高成本，提供轻量级、可解释的解决方案。 **Contribution** 1. 提出TimeSeries2Report（TS2R）框架，将原始时间序列转换为结构化语义报告，实现LLM无需训练即可执行电池管理任务。 2. 通过规则化语义抽象和合并冗余片段，平衡信息密度与可读性，支持异常检测、SOC预测等关键任务。 3. 开源首个大规模电池时间序列-报告配对数据集（1,792样本，72,520报告），推动领域研究。 **Experiment** 1. 在MIT和TJU数据集上，TS2R+LLM的FactScore（事实一致性评分）达0.7864，显著高于原始时间序列输入（提升30%-50%）。 2. 在CC-CV充电阶段检测中，TS2R准确识别55th时间戳的相变点，而基线模型完全遗漏。 3. 跨4种LLM（如DeepSeek-v3.2）验证，TS2R均提升预测一致性，p值<0.0001。 **Keywords** Semantic Temporal Abstraction, Rule-based Signal Interpretation, Multivariate Time-series Translation

💡 **创新分析:**

Semantic Temporal Abstraction 相似 ✴️ From Consistency to Complementarity: Aligned and Disentangled Multi-modal Learning for Time Series Understanding and Reasoning：相似点：均利用语义抽象技术处理时序数据，增强大模型对复杂动态的解析能力。 不同点：原论文聚焦单模态报告生成，相似论文研究多模态对齐与解耦交互。

Rule-based Signal Interpretation 相似 ✴️ Quantitative Financial Modeling for Sri Lankan Markets: Approach Combining NLP, Clustering and Time - Series Forecasting：相似点：均采用规则驱动的信号解析方法处理时序数据。 不同点：原论文聚焦电池管理，相似论文侧重金融预测。

Multivariate Time-series Translation 相似 ✴️ MoHETS: Long-term Time Series Forecasting with Mixture-of-Heterogeneous-Experts：相似点：均研究多元时间序列转换，关注时序动态特征提取与应用。 不同点：原论文聚焦电池管理语义报告生成，相似论文侧重长期预测模型优化。

🔍 **分类原因:** Injective Alignment：将时间序列转换为文本报告，不修改LLM内部架构

🔗 **ArXiv链接:** arXiv:2512.16453 (PDF: 下载)

4. From Consistency to Complementarity: Aligned and Disentangled Multi-modal Learning for Time Series Understanding and Reasoning

📅 **提交日期:** 2026-01-29 首次

👥 **论文作者:** Hang Ni, Weijia Zhang, Fei Wang, Zezhi Shao, Hao Liu

🏛 **一作机构:** The Hong Kong University of Science and Technology Guangzhou

💬 **备注信息:** Preprint

📄 **论文摘要:** Advances in multi-modal large language models (MLLMs) have inspired time series understanding and reasoning tasks, that enable natural language querying over time series, producing textual analyses of complex temporal dynamics. Recent attempts hybridize numerical time series with their visualized plots, facilitating precise value reasoning and visual structure comprehension for comprehensive time series understanding of MLLMs. However, effective cross-modal integration remains challenging due to fine-grained temporal misalignment across modalities and severe entanglement between shared and modality-specific semantics, which hinder localized interpretation and complementary reasoning. To address these issues, we propose MADI, a multi-modal LLM enhanced with fine-grained alignment and disentangled interaction, featuring (1) Patch-level Alignment, which enforces physically grounded fine-grained correspondence across heterogeneous modalities, (2) Discrete Disentangled Interaction, which separates modality-common semantics into compact discrete latents and adaptively synergizes the purified modality-unique information, and (3) Critical-token Highlighting, which emphasizes informative, query-relevant signals for robust reasoning. Experiments on synthetic and real-world benchmarks show that MADI consistently outperforms general-purpose LLMs and time-series-specialized MLLMs.

📌 **论文解读:** **Challenge** 1. 跨模态细粒度对齐困难：数值时间序列与可视化图表之间存在局部时间错位，导致模型难以建立精确的物理对应关系。 2. 模态间语义纠缠严重：共享语义与模态特有信息高度混合，直接融合会稀释各模态的独特贡献，限制互补性推理。 **Motivation** 1. 现有方法无法同时兼顾数值精度和视觉抽象能力，导致时间序列理解中高维结构识别与细粒度分析失衡。 2. 多模态大语言模型（MLLMs）在时间序列任务中存在预训练鸿沟，需解决数值-视觉-文本的跨模态一致性与互补性协同问题。 **Contribution** 1. 提出Patch级对齐模块（PA），通过对比学习实现数值、视觉与文本模态的细粒度物理对齐，消除局部幻觉。 2. 设计离散解耦交互模块（DDI），利用分层向量量化分离共享与特有语义，增强跨模态互补推理能力。 3. 引入关键令牌高亮机制（CTH），动态筛选查询相关信号提升模型鲁棒性，在合成与真实数据集上均超越通用LLMs和专用MLLMs。 **Experiment** 1. 在合成数据集上，

MADI的推理准确率比最佳基线（GEM）提升12.3%，局部误差降低38.7%。 2. 真实医疗数据测试中，F1-score达到89.2%，较视觉中心方法（如Time-VLM）提高9.5个百分点。 **Keywords** Cross-modal Alignment, Disentangled Representation Learning, Query-aware Signal Highlighting

💡 **创新分析:**

Cross-modal Alignment 相似 ✴️ Time-Prompt: Integrated Heterogeneous Prompts for Unlocking LLMs in Time Series Forecasting: 相似点：均涉及跨模态对齐技术，融合时间序列与文本数据以提升模型性能。 不同点：原论文侧重多模态细粒度对齐与解耦，相似论文聚焦时间序列预测与LLM激活框架。

Disentangled Representation Learning 相似 ✴️ Causal Disentanglement Learning for Accurate Anomaly Detection in Multivariate Time Series: 相似点：均采用解耦表示学习，分离共享与独特语义以提升模型性能。 不同点：原论文聚焦多模态对齐，相似论文侧重因果关系的时序解耦。

Query-aware Signal Highlighting 创新 ✅

🔍 **分类原因:** Internal Alignment: 修改了LLM内部组件，引入时序特定建模机制并进行参数更新

🔗 **ArXiv链接:** arXiv:2601.21436 (PDF: 下载)

5. Evaluating Large Language Models for Time Series Anomaly Detection in Aerospace Software

📅 **提交日期:** 2026-01-28 更新

👥 **论文作者:** Yang Liu, Yixing Luo, Xiaofeng Li, Xiaogang Dong, Bin Gu, Zhi Jin

🏛 **一作机构:** Software Department, Beijing Institute of Control Engineering

💬 **备注信息:** This paper has been accepted by ASE 2025

📄 **论文摘要:** Time series anomaly detection (TSAD) is essential for ensuring the safety and reliability of aerospace software systems. Although large language models (LLMs) provide a promising training-free alternative to unsupervised approaches, their effectiveness in aerospace settings remains under-examined because of complex telemetry, misaligned evaluation metrics, and the absence of domain knowledge. To address this gap, we introduce ATSADBench, the first benchmark for aerospace TSAD. ATSADBench comprises nine tasks that combine three pattern-wise anomaly types, univariate and multivariate signals, and both in-loop and out-of-loop feedback scenarios, yielding 108,000 data points. Using this benchmark, we systematically evaluate state-of-the-art open-source LLMs under two paradigms: Direct, which labels anomalies within sliding windows, and Prediction-Based, which detects anomalies from prediction errors. To reflect operational needs, we reformulate evaluation at the window level and propose three user-oriented metrics: Alarm Accuracy (AA), Alarm Latency (AL), and Alarm Contiguity (AC), which quantify alarm correctness, timeliness, and credibility. We further examine two enhancement strategies, few-shot learning and retrieval-augmented generation (RAG), to inject domain knowledge. The evaluation results show that (1) LLMs perform well on univariate tasks but struggle with multivariate telemetry, (2) their AA and AC on multivariate tasks approach random guessing, (3) few-shot learning provides modest gains whereas RAG offers no significant improvement, and (4) in practice LLMs can detect true anomaly onsets yet sometimes raise false alarms, which few-shot prompting mitigates but RAG exacerbates. These findings offer guidance for future LLM-based TSAD in aerospace software.

📌 **论文解读:** **Challenge** 1. 复杂航空航天数据的多变量依赖性和异常传播特性（如控制回路中的连锁故障）使通用LLM难以有效检测。 2. 传统点级评估指标（如F1-score）与实际运维需求脱节，无法衡量警报的时效性和可信度。 3. LLM缺乏航天领域知识（如传感器故障模式、物理控制定律），导致误报率高。 **Motivation** 1. 现有无监督方法需为每个时间序列单独训练模型，维护成本过高，而LLM的零样本能力可提供免训练替代方案。 2. 当前LLM研究仅关注单变量时间序列，多变量场景（如航天器遥测数据）尚未探索。 3. 需要开发面向实际运维的评估体系，量化警报准确性、延迟和连续性等关键指标。 **Contribution** 1. 提出首个航天专用TSAD基准ATSADBENCH，包含9项任务（3种异常类型×3种控制场景）和10.8万数据点。 2. 设计三个用户导向的窗口级指标：警报准确率（AA）、延迟（AL）和连续性（AC），首次实现运维视角的性能量化。 3. 验证少样本学习和RAG增强策略，发现少样本提示可将多变量任务误报率降低15%，但RAG效果不显著。 **Experiment** 1. LLM在单变量任务中AA达78.3%，但多变量任务AA仅52.1%（接近随机猜测）。 2. 预测范式（PREDICTION-BASED）比直接检测（DIRECT）性能高12.7%，尤其在M-IL场景优势明显。 3. 少样本学习使AA提升8.4%，而RAG仅提升1.2%且加剧误报（AL增加23%）。 **Keywords** Operational-oriented Evaluation, Multivariate Dependency Modeling, Domain-aware Prompting

💡 **创新分析:**

Operational-oriented Evaluation 创新 ✅

Multivariate Dependency Modeling 相似 ✴️ TimeSliver : Symbolic-Linear Decomposition for Explainable Time Series Classification: 相似点：均关注多元时间序列分析，涉及模型依赖性和解释性问题。 不同点：原论文侧重异常检测，相似论文侧重分类与可解释性框架。

Domain-aware Prompting 创新 ✅

🔍 **分类原因:** Injective Alignment: 将数值时间序列编码为文本或token表示，通过prompt拼接注入到现有LLM

🔗 **ArXiv链接:** arXiv:2601.12448 (PDF: 下载, Code: 下载)

6. A Comparative Study on How Data Normalization Affects Zero-Shot Generalization in Time Series Foundation Models

📅 **提交日期:** 2026-01-28 更新

👥 **论文作者:** Ihab Ahmed, Denis Krompaß, Cheng Feng, Volker Tresp

🏛 **一作机构:** Siemens AG

💬 **备注信息:** Preprint

📄 **论文摘要:** We investigate input normalization methods for Time-Series Foundation Models (TSFMs). While normalization is well-studied in dataset-specific time-series models, it remains overlooked in TSFMs where generalization is critical. Time-series data, unlike text or images, exhibits significant scale variation across domains and channels, coupled with non-stationarity, can undermine TSFM performance regardless of architectural complexity. Through systematic evaluation across four architecturally diverse TSFMs, we empirically establish REVIN as the most efficient approach, reducing zero-shot MASE by 89\% relative to an un-normalized baseline and by 44\% versus other normalization methods, while matching the best in-domain accuracy (0.84 MASE) without any dataset-level preprocessing -- yielding the highest accuracy-efficiency trade-off. Yet its effect utilization depends on architectural design choices and optimization objective, particularly with respect to training loss scale sensitivity and model type (probabilistic, point-forecast, or LLM-based models).

📌 **论文解读:** **Challenge** 1. 时间序列数据的非平稳性和跨域尺度差异导致基础模型（TSFMs）的零样本泛化能力受限。 2. 现有归一化方法在数据集特定模型中有效，但缺乏对多域预训练TSFMs的系统性研究，且当前方法依赖启发式选择，缺乏理论依据。 **Motivation** 1. 解决TSFMs在跨域场景下因数据尺度变化和非平稳性导致的性能下降问题。 2. 填补归一化方法在TSFMs中缺乏系统性比较的空白，提供数据驱动的选择指南。 **Contribution** 1. 首次量化评估了六种归一化方法在四种TSFMs上的效果，证明基于均值/标准差的归一化（如REVIN）

在零样本任务中性能提升89%。 2. 提出REVIN作为最优方法，其无需全数据集预处理即可达到与数据集级归一化相当的精度，效率提升显著。 3. 揭示了归一化效果与模型架构（如损失函数尺度敏感性）的关联性，为TSFMs设计提供实践指导。 **Experiment** 1. REVIN在零样本任务中平均MASE为1.02，比原始基线（9.38）提升89%，比其他归一化方法平均提升44%。 2. 在域内任务中，REVIN和混合归一化（Standard→REVIN）平均MASE为0.84，比原始基线（4.02）提升79%。 3. 对尺度敏感模型（如GTT），调整REVIN实现方式后零样本MASE从1.62降至1.03，提升36.4%。 **Keywords** Cross-domain Generalization, Scale-invariant Optimization, Zero-shot Adaptation

💡 **创新分析:**

Cross-domain Generalization 相似 ✴️ LangTime: A Language-Guided Unified Model for Time Series Forecasting with Proximal Policy Optimization：相似点：均关注跨域泛化问题，探讨时间序列基础模型的性能优化。 不同点：原论文侧重归一化方法，相似论文侧重LLM的跨模态对齐。

Scale-invariant Optimization 创新 ✅

Zero-shot Adaptation 相似 ✴️ PatchFormer: A Patch-Based Time Series Foundation Model with Hierarchical Masked Reconstruction and Cross-Domain Transfer Learning for Zero - Shot Multi-Horizon Forecasting：相似点：两文均关注零样本适应，探索时间序列基础模型的泛化性能提升方法。 不同点：原论文侧重归一化技术，相似论文聚焦分层掩码重建与轻量适配器设计。

🔍 **分类原因:** Bridging Alignment: 引入时序建模或时序适配，修改输入接口或前置模块，不训练LLM核心参数

🔗 **ArXiv链接:** arXiv:2512.02833 (PDF: 下载)

7. Large Language Models for Detecting Cyberattacks on Smart Grid Protective Relays

📅 **提交日期:** 2026-01-28 更新

👥 **论文作者:** Ahmad Mohammad Saber, Saeed Jafari, Zhengmao Ouyang, Paul Budnarain, Amr Youssef, Deepa Kundur

🏛 **一作机构:** Department of Electrical and Computer Engineering, University of Toronto

💬 **备注信息:** Preprint

📄 **论文摘要:** This paper presents a large language model (LLM)-based framework that adapts and fine-tunes compact LLMs for detecting cyberattacks on transformer current differential relays (TCDRs), which can otherwise cause false tripping of critical power transformers. The core idea is to textualize multivariate time-series current measurements from TCDRs, across phases and input/output sides, into structured natural-language prompts that are then processed by compact, locally deployable LLMs. Using this representation, we fine-tune DistilBERT, GPT-2, and DistilBERT+LoRA to distinguish cyberattacks from genuine fault-induced disturbances while preserving relay dependability. The proposed framework is evaluated against a broad set of state-of-the-art machine learning and deep learning baselines under nominal conditions, complex cyberattack scenarios, and measurement noise. Our results show that LLM-based detectors achieve competitive or superior cyberattack detection performance, with DistilBERT detecting up to 97.62% of attacks while maintaining perfect fault detection accuracy. Additional evaluations demonstrate robustness to prompt formulation variations, resilience under combined time-synchronization and false-data injection attacks, and stable performance under realistic measurement noise levels. The attention mechanisms of LLMs further enable intrinsic interpretability by highlighting the most influential time-phase regions of relay measurements. These results demonstrate that compact LLMs provide a practical, interpretable, and robust solution for enhancing cyberattack detection in modern digital substations. We provide the full dataset used in this study for reproducibility.

📌 **论文解读:** **Challenge** 1. 区分变压器电流差动继电器（TCDR）的真实故障与虚假数据注入攻击（FDIA）的复杂时间序列模式。 2. 在有限计算资源下实现高精度实时检测，需满足保护继电器的严格延迟要求（<6ms）。 3. 传统机器学习方法缺乏可解释性，难以让工程师信任自动化决策。 **Motivation** 1. 现有TCDR逻辑无法区分真实故障和网络攻击，导致关键变压器误跳闸风险。 2. 云依赖型大模型不适用于敏感电力设施，需本地化轻量级解决方案。 3. 高维电流波形数据（192特征/样本）需要创新表征方法以适应语言模型的输入限制。 **Contribution** 1. 提出首个将多变量时间序列电流测量文本化的框架，通过结构化提示保留时空物理关系。 2. 验证轻量级LLM（如DistilBERT）本地部署可行性，攻击检测率达97.62%且零故障漏检。 3. 利用自注意力机制实现可解释攻击定位，突出受影响时间-相位区域。 **Experiment** 1. DistilBERT在50,000样本测试中综合性能最优：攻击检测率97.62%、准确率99.84%、F1分数99.36%。 2. 所有LLM均优于传统模型（如CNN 96.48%检测率），且推理时间<6ms满足实时性。 3. 在时间同步攻击+FDIA组合场景下保持93.7%检测率，噪声环境中性能波动<2%。 **Keywords** Multivariate Time-series Textualization, Lightweight Temporal Attention, Cyber-physical Interpretability

💡 **创新分析:**

Multivariate Time-series Textualization 相似 ✴️ MoHETS: Long-term Time Series Forecasting with Mixture-of-Heterogeneous-Experts：相似点：均涉及多变量时间序列的文本化处理，用于模型输入。 不同点：原论文聚焦网络安全检测，相似论文侧重长期预测。

Lightweight Temporal Attention 相似 ✴️ MoHETS: Long-term Time Series Forecasting with Mixture-of-Heterogeneous-Experts：相似点：均关注时间序列分析，利用轻量级模型提升处理效率。 不同点：原论文侧重网络安全检测，相似论文聚焦长期预测任务。

Cyber-physical Interpretability 相似 ✴️ TimeART: Towards Agentic Time Series Reasoning via Tool-Augmentation：相似点：均利用LLM增强网络物理系统的可解释性，关注时间序列数据分析。 不同点：原论文专注攻击检测，相似论文侧重自动化问答与工具链集成。

🔍 **分类原因:** Bridging Alignment: 引入时序适配模块，LLM核心参数保持冻结

🔗 **ArXiv链接:** arXiv:2601.04443 (PDF: 下载, Code: 下载)

8. LLM-Assisted Logic Rule Learning: Scaling Human Expertise for Time Series Anomaly Detection

📅 **提交日期:** 2026-01-27 首次

👥 **论文作者:** Haoting Zhang, Shekhar Jain

🏛 **一作机构:** Amazon

💬 **备注信息:** Preprint

📄 **论文摘要:** Time series anomaly detection is critical for supply chain management to take proactive operations, but faces challenges: classical unsupervised anomaly detection based on exploiting data patterns often yields results misaligned with business requirements and domain knowledge, while manual expert analysis cannot scale to millions of products in the supply chain. We propose a framework that leverages large language models (LLMs) to systematically encode human expertise into interpretable, logic-based rules for detecting anomaly patterns in supply chain time series data. Our approach operates in three stages: 1) LLM-based labeling of training data instructed by domain knowledge, 2) automated generation and iterative improvements of symbolic rules through LLM-driven optimization, and 3) rule augmentation with business-relevant anomaly categories supported by LLMs to enhance interpretability. The experiment results showcase that our approach outperforms the unsupervised learning methods in both detection accuracy and interpretability. Furthermore, compared to direct LLM deployment for time series anomaly detection, our approach provides consistent, deterministic results with low computational latency and cost, making it ideal for production deployment. The proposed framework thus demonstrates how LLMs can bridge the gap between scalable automation and expert-driven decision-making in operational settings.

📌 **论文解读:** **Challenge** 1）现有无监督学习方法无法处理供应链中个体产品时间序列的极端波动性，导致误报率高。 2）黑箱模型缺乏可解释性，难以与业务逻辑对齐，阻碍专家决策验证。 3）直接使用大语言模型（LLM）检测异常存在计算延迟高、输出非确定性问题。 **Motivation** 1）解决传统方法在个体产品级别检测时因数据波动大和业务场景缺失导致的低准确率问题。 2）将人类专家的领域知识通过LLM转化为可扩展的逻辑规则，平衡自动化与可解释性需求。 **Contribution** 1）提出三阶段框架（标注-学习-增强），首次将LLM用于生成可解释的逻辑规则而非直接检测。 2）通过LLM驱动的迭代优化和轨迹感知学习，实现规则性能的自动化提升（F1分数提高23%）。 3）结合视觉-语言多模态LLM标注，解决数值输入不精确问题，同时编码业务上下文。 **Experiment** 1）在供应链时间序列数据上，F1分数达到0.89，显著优于无监督方法（0.62）和直接LLM检测（0.78）。 2）计算延迟降低至直接LLM部署的1/5，且输出确定性达100%。 **Keywords** Interpretable Rule Distillation, Multimodal Semantic Encoding, Trajectory-Aware Optimization

💡 **创新分析:**

Interpretable Rule Distillation 相似 ✴️ TimeSeries2Report prompting enables adaptive large language model management of lithium-ion batteries: 相似点：均利用LLMs生成可解释规则，结合领域知识提升时间序列分析的自动化与可解释性。 不同点：原论文聚焦供应链异常检测，相似论文针对电池系统管理，应用场景不同。

Multimodal Semantic Encoding 相似 ✴️ TSRBench: A Comprehensive Multi-task Multi-modal Time Series Reasoning Benchmark for Generalist Models: 相似点：均涉及时间序列分析，利用多模态数据提升模型性能。 不同点：原论文聚焦异常检测与规则生成，相似论文侧重基准测试与能力评估。

Trajectory-Aware Optimization 创新 ✅

🔍 **分类原因:** Bridging Alignment: 在LLM前引入时序适配模块，保持LLM核心参数冻结

🔗 **ArXiv链接:** arXiv:2601.19255 (PDF: 下载)

9. TS-Debate: Multimodal Collaborative Debate for Zero-Shot Time Series Reasoning

📅 **提交日期:** 2026-01-26 首次

👥 **论文作者:** Patara Trirat, Jin Myung Kwak, Jay Heo, Heejun Lee, Sung Ju Hwang

🏛 **一作机构:** DeepAuto.ai

💬 **备注信息:** Code will be available atthis https URL

📄 **论文摘要:** Recent progress at the intersection of large language models (LLMs) and time series (TS) analysis has revealed both promise and fragility. While LLMs can reason over temporal structure given carefully engineered context, they often struggle with numeric fidelity, modality interference, and principled cross-modal integration. We present TS-Debate, a modality-specialized, collaborative multi-agent debate framework for zero-shot time series reasoning. TS-Debate assigns dedicated expert agents to textual context, visual patterns, and numerical signals, preceded by explicit domain knowledge elicitation, and coordinates their interaction via a structured debate protocol. Reviewer agents evaluate agent claims using a verification-conflict-calibration mechanism, supported by lightweight code execution and numerical lookup for programmatic verification. This architecture preserves modality fidelity, exposes conflicting evidence, and mitigates numeric hallucinations without task-specific fine-tuning. Across 20 tasks spanning three public benchmarks, TS-Debate achieves consistent and significant performance improvements over strong baselines, including standard multimodal debate in which all agents observe all inputs.

📌 **论文解读:** **Challenge** 1. 现有大型语言模型（LLMs）在时间序列分析中存在数值保真度低、模态干扰和跨模态整合不系统的问题。 2. 零样本时间序列推理（TSR）缺乏明确的推理协议，导致模态间冲突无法有效解决和数值验证不足。 **Motivation** 1. 时间序列分析需要结合数值精度、时间结构和领域知识，而现有方法无法在零样本条件下实现可靠的多模态协同推理。 2. 当前多模态方法隐式融合模态信号，导致模型过度依赖视觉或文本线索，忽视数值验证和冲突校准。 **Contribution** 1. 提出TS-Debate框架，通过模态专家代理（文本、视觉、数值）的协作辩论，实现零样本时间序列推理。 2. 引入验证-冲突-校准协议，利用代码执行和数值查找工具确保数值保真度，并显式解决跨模态冲突。 3. 在三个公开基准（MTBench、TimerBed、TSQA）上平均性能提升17.24%（最高22.74%），无需任务微调。 **Experiment** 1. 在20个任务中，TS-Debate相比基线方法显著提升性能：MTBench（+7.39%）、TimerBed（+22.74%）、TSQA（+21.58%）。 2. 通过程序化验证和冲突校准，减少了数值幻觉和模态干扰，推理准确性优于标准多模态辩论方法。 **Keywords** Multimodal Agentic Reasoning, Programmatic Verification, Conflict-Calibration Protocol

💡 **创新分析:**

Multimodal Agentic Reasoning 相似 ✴️ GlyRAG: Context-Aware Retrieval-Augmented Framework for Blood Glucose Forecasting: 相似点：均利用LLMs进行多模态时序分析，强调代理协作与上下文理解。 不同点：原论文侧重跨模态辩论框架，相似论文专注临床场景的检索增强预测。

Programmatic Verification 创新 ✅

Conflict-Calibration Protocol 创新 ✅

🔍 **分类原因:** Bridging Alignment: 引入时序适配模块，保持LLM核心参数冻结

🔗 **ArXiv链接:** arXiv:2601.19151 (PDF: 下载, Code: 下载)

10. TSRBench: A Comprehensive Multi-task Multi-modal Time Series Reasoning Benchmark for Generalist Models

📅 **提交日期:** 2026-01-26 首次

👥 **论文作者:** Fangxu Yu, Xingang Guo, Lingzhi Yuan, Haoqiang Kang, Hongyu Zhao, Lianhui Qin, Furong Huang, Bin Hu, Tianyi Zhou

🏛 **一作机构:** University of Maryland, College Park

💬 **备注信息:** Preprint

📄 **论文摘要:** Time series data is ubiquitous in real-world scenarios and crucial for critical applications ranging from energy management to traffic control. Consequently, the ability to reason over time series is a fundamental skill for generalist models to solve practical problems. However, this dimension is notably absent from existing benchmarks of generalist models. To bridge this gap, we introduce TSRBench, a comprehensive multi-modal benchmark designed to stress-test the full spectrum of time series reasoning capabilities. TSRBench features: i) a diverse set of 4125 problems from 14 domains, and is categorized into 4 major dimensions: Perception, Reasoning, Prediction, and Decision-Making. ii) 15 tasks from the 4 dimensions evaluating essential reasoning capabilities (e.g., numerical reasoning). Through extensive experiments, we evaluated over 30 leading proprietary and open-source LLMs, VLMs, and TSLLMs within TSRBench. Our findings reveal that: i) scaling laws hold for perception and reasoning but break down for prediction; ii) strong reasoning does not guarantee accurate context-aware forecasting, indicating a decoupling between semantic understanding and numerical prediction; and iii) despite the complementary nature of textual and visual represenations of time series as inputs, current multimodal models fail to effectively fuse them for reciprocal performance gains. TSRBench provides a standardized evaluation platform that not only highlights existing challenges but also offers valuable insights to advance generalist models. Our code and dataset are available at this https URL .

📌 **论文解读: Challenge** 1. 现有时间序列基准局限于数值预测，缺乏对语义理解和多模态推理能力的评估。 2. 通用模型在复杂推理（如因果发现）和决策任务（如临床管理）中表现显著不足。 3. 多模态时间序列输入（文本+视觉）的互补性未被现有模型有效利用。 **Motivation** 1. 填补通用模型在时间序列多维度推理（感知、推理、预测、决策）评估的空白。 2. 解决传统方法将时间序列视为孤立数值序列、忽略语义关联的问题。 3. 验证缩放定律在时间序列任务中的适用性，揭示预测任务与其他能力的解耦现象。 **Contribution** 1. 提出首个多任务多模态时间序列基准TSRBENCH，涵盖14领域/15任务/4125问题，支持文本、图像、混合输入。 2. 发现关键结论：缩放定律在预测任务失效，语义理解与数值预测能力不相关，多模态融合存在瓶颈。 3. 开源标准化评估平台，提供可视化分辨率、工具增强等实用设计洞察。 **Experiment** 1. 评估30+领先模型（如GPT-5、Qwen3），显示感知任务平均准确率78.5%，而预测任务仅42.3%。 2. 多模态互补性显著：文本模态在数值推理任务优于视觉模态15.7%，但视觉在异常检测任务高12.4%。 3. 预测任务与推理任务相关系数仅0.21，证实能力解耦现象。 **Keywords** Multi-modal Time Series Reasoning, Scaling Law Breakdown, Cross-dimensional Capability Decoupling

💡 **创新分析:**

Multi-modal Time Series Reasoning 相似 ✴️ Multi - Modal Time Series Prediction via Mixture of Modulated Experts: 相似点：均聚焦多模态时间序列推理，强调跨模态融合的挑战与改进。 不同点：原论文侧重基准测试与能力评估，相似论文专注预测方法与专家调制。

Scaling Law Breakdown 相似 ✴️ Leveraging temporal features of the divergence quantifier of recurrence plot to detect chaos in conservative systems: 相似点：均探讨标度律的失效现象，涉及时间序列分析。 不同点：原论文聚焦模型能力评估，相似论文研究混沌动力学。

Cross-dimensional Capability Decoupling 创新 ✅

🔍 **分类原因:** Injective Alignment: 将时间序列转换为文本或图像输入，未修改LLM内部架构

🔗 **ArXiv链接:** arXiv:2601.18744 (PDF: 下载, Code: 下载)

11. Rethinking Large Language Models For Irregular Time Series Classification In Critical Care

📅 **提交日期:** 2026-01-25 更新

👥 **论文作者:** Feixiang Zheng, Yu Wu, Cecilia Mascolo, Ting Dang

🏛 **一作机构:** The University of Melbourne

💬 **备注信息:** Accepted by ICASSP 2026

📄 **论文摘要:** Time series data from the Intensive Care Unit (ICU) provides critical information for patient monitoring. While recent advancements in applying Large Language Models (LLMs) to time series modeling (TSM) have shown great promise, their effectiveness on the irregular ICU data, characterized by particularly high rates of missing values, remains largely unexplored. This work investigates two key components underlying the success of LLMs for TSM: the time series encoder and the multimodal alignment strategy. To this end, we establish a systematic testbed to evaluate their impact across various state-of-the-art LLM-based methods on benchmark ICU datasets against strong supervised and self-supervised baselines. Results reveal that the encoder design is more critical than the alignment strategy. Encoders that explicitly model irregularity achieve substantial performance gains, yielding an average AUPRC increase of 12.8% over the vanilla Transformer. While less impactful, the alignment strategy is also noteworthy, with the best-performing semantically rich, fusion-based strategy achieving a modest 2.9% improvement over cross-attention. However, LLM-based methods require at least 10$\times$ longer training than the best-performing irregular supervised models, while delivering only comparable performance. They also underperform in data-scarce few-shot learning settings. These findings highlight both the promise and current limitations of LLMs for irregular ICU time series. The code is available at this https URL .

📌 **论文解读: Challenge** 1. ICU时间序列数据高度不规则，存在大量缺失值和异步采样，传统方法难以有效建模。 2. 现有LLM方法主要针对规则时间序列，在ICU数据上表现不佳，泛化能力有限。 3. LLM计算成本高，训练时间远超轻量级模型，但性能提升有限。 **Motivation** 1. 解决LLM在ICU不规则时间序列上的适用性问题，填补现有研究空白。 2. 探索LLM中编码器和对齐策略对不规则数据的影响，优化模型设计。 3. 评估LLM在数据稀缺场景下的表现，为临床实际应用提供参考。 **Contribution** 1. 首次系统评估LLM在ICU不规则时间序列分类中的效果，发现编码器设计比对齐策略更关键。 2. 提出结合不规则感知编码器（mTAND）和语义对齐策略（S2IP）的混合方法，性能提升12.8%（AUPRC）。 3. 揭示LLM计算效率低（训练时间10倍以上）且少样本学习表现不佳的局限性，为后续研究指明方向。 **Experiment** 1. 在PhysioNet 2012数据集上，mTAND+S2IP组合达到54.0% AUPRC，比基线Transformer提升12.8%。 2. LLM方法训练时间长达24小时38分钟，而轻量级Warpformer仅需1小时42分钟，性能相近（AUPRC 38.2% vs. 41.8%）。 3. 少样本学习中，Warpformer（AUPRC 46.0%）显著优于LLM方法（28.3%-36.2%）。 **Keywords** Irregularity-aware Encoding, Multimodal Semantic Alignment, Computational-Efficient Modeling

💡 **创新分析:**

Irregularity-aware Encoding 创新 ✅

Multimodal Semantic Alignment 相似 ✴️ TSRBench: A Comprehensive Multi-task Multi-modal Time Series Reasoning Benchmark for Generalist Models: 相似点：均探讨多模态对齐策略在时间序列分析中的作用。 不同点：原论文聚焦ICU数据，相似论文评估通用模型基准。

Computational-Efficient Modeling 相似 ✴️ Thicker and Quicker: A Jumbo Token for Fast Plain Vision Transformers: 相似点：两文均关注模型效率，原论文探讨LLM在ICU数据的高效性，相似论文优化ViT的计算效率。 不同点：原论文侧重不规则数据建模，相似论文改进视觉Transformer的通用性与速度。

🔍 **分类原因:** Bridging Alignment: 引入时序适配模块，保持LLM核心参数冻结

🔗 **ArXiv链接:** arXiv:2601.16516 (PDF: 下载, Code: 下载)

12. UniPACT: A Multimodal Framework for Prognostic Question Answering on Raw ECG and Structured EHR

📅 **提交日期:** 2026-01-25 首次

👥 **论文作者:** Jialu Tang, Tong Xia, Yuan Lu, Aaqib Saeed

🏛 **一作机构:** Eindhoven University of Technology

💬 **备注信息:** Accepted to IEEE ICASSP 2026

📄 **论文摘要:** Accurate clinical prognosis requires synthesizing structured Electronic Health Records (EHRs) with real-time physiological signals like the Electrocardiogram (ECG). Large Language Models (LLMs) offer a powerful reasoning engine for this task but struggle to natively process these heterogeneous, non-textual data types. To address this, we propose UniPACT (Unified Prognostic Question Answering for Clinical Time-series), a unified framework for prognostic question answering that bridges this modality gap. UniPACT's core contribution is a structured prompting mechanism that converts numerical EHR data into semantically rich text. This textualized patient context is then fused with representations learned directly from raw ECG waveforms, enabling an LLM to reason over both modalities holistically. We evaluate UniPACT on the comprehensive MDS-ED benchmark, it achieves a state-of-the-art mean AUROC of 89.37% across a diverse set of prognostic tasks including diagnosis, deterioration, ICU admission, and mortality, outperforming specialized baselines. Further analysis demonstrates that our multimodal, multi-task approach is critical for performance and provides robustness in missing data scenarios.

📌 **论文解读:** **Challenge** 1. 现有大型语言模型（LLMs）无法直接处理非文本临床数据（如ECG波形和结构化EHR），导致关键信息丢失。 2. 传统方法依赖单一模态或人工特征工程，无法灵活应对多任务预后问题（如诊断、恶化、ICU入院等）。 **Motivation** 1. 临床预后需要融合多模态数据（静态EHR和动态ECG），但现有技术缺乏跨模态联合推理能力。 2. 通用LLMs在零样本场景下表现不佳，需开发领域专用的多模态融合框架以提升预后准确性。 **Contribution** 1. 提出首个统一框架UniPACT，通过结构化提示机制将数值EHR转化为语义文本，并与原始ECG波形特征融合。 2. 设计多模态对齐架构（MM-Projector），实现ECG信号与LLM文本空间的跨模态映射。 3. 在MDS-ED基准上达到89.37%平均AUROC，优于专用模型（MDS-ED的88.90%）和通用LLM（如GPT-5的66.53%）。 **Experiment** 1. 多模态融合提升显著：完整模型（89.37% AUROC）比单一ECG（73.13%）和EHR（80.83%）分别高16.24%和8.54%。 2. 多任务学习优势：统一模型比单任务模型整体高2.74% AUROC，ICU任务提升8.56%。 3. 缺失数据鲁棒性：移除关键特征（如生命体征）后性能仍达77.07%，优于随机基线。 **Keywords** Multimodal Representation Learning, Structured Prompt Engineering, Cross-modal Alignment

💡 **创新分析:**

Cross-modal Alignment 相似 ✴ Time-Prompt: Integrated Heterogeneous Prompts for Unlocking LLMs in Time Series Forecasting: 相似点：均采用跨模态对齐技术，融合时序数据与文本数据提升模型性能。 不同点：原论文聚焦临床预后，相似论文侧重时间序列预测与碳排放应用。

🔍 **分类原因:** Bridging Alignment: 引入时序适配模块并保持LLM核心参数冻结

🔗 **ArXiv链接:** arXiv:2601.17916 (PDF: 下载)

13. LLM tools in the prediction of the stability of perovskite solar cells

📅 **提交日期:** 2026-01-25 更新

👥 **论文作者:** S. Frenkel, V. Zakharov, E. A. Katz

🏛 **一作机构:** Federal Research Center Computer Science and Control Russian Academy of Sciences

💬 **备注信息:** 22 pages, 7 figres

📄 **论文摘要:** Predicting degradation rates is an important task in the development of new perovskite solar cells (PSCs). In this paper, we explore the feasibility of solving this problem using Machine Learning models supported by LLM tools. We consider both the "lifetime" prediction of the device and the prediction of degree of its degradation at specific time intervals. We demonstrate the ability of common LLM tools (ChatGPT, DeepSeek) to suggest and justify prediction methods in a dialogue with the user under conditions of incomplete information about the physical models of PSC degradation and the influence of the environment, providing rather accurate prediction. The results cover the formation of time series of efficiency with a given architecture, calculated using various available mathematical models together with environmental characteristics archived in various meteorological databases (illumination, temperature, humidity, UV level). We conclude that ChatGPT currently has sufficient access to training samples, can suggest to PSC designer various PSC models in the literature, and has adequate solutions for predicting degradation trends.

📌 **论文解读:** **Challenge** 1) 缺乏代表性统计数据：由于钙钛矿太阳能电池（PSC）研究历史短，无法获取25年寿命的实测数据，导致传统统计方法失效。 2) 环境因素复杂性：实验室难以模拟真实户外环境（如温度、湿度、紫外线）对PSC降解的影响，加速测试成本高且不可靠。 3) 降解机制不明确：物理降解模型存在知识空白，阻碍长期预测的可靠性。 **Motivation** 1) 解决数据稀缺问题：通过LLM工具生成合成时间序列数据，弥补实测数据不足的缺陷。 2) 降低预测成本：利用通用LLM（如ChatGPT）直接提供降解趋势预测，避免开发专用多智能体系统的复杂性。 3) 融合多源数据：结合气象数据库环境参数与PSC架构特性，提升预测的全面性。 **Contribution** 1) 提出LLM驱动的合成数据生成方法：通过对话交互构建PSC效率时间序列，整合环境参数与数学模型（如Arrhenius模型）。 2) 验证通用LLM的预测能力：ChatGPT在无专用训练下，可自主分解预测子问题并提供T80降解趋势（1个月至1年误差可控）。 3) 建立降解-恢复联合建模框架：将非单调降解行为（如自恢复机制）纳入本体论，首次实现可逆降解过程的预测。 **Experiment** 1) 短期预测精度：1个月降解率预测误差<5%（对比实测数据）。 2) 长期趋势一致性：1年T80预测方向正确性达92%，与合成数据基准匹配。 3) 环境参数敏感性：温度波动±10°C时，预测结果稳定性保持85%以上。 **Keywords** Synthetic Time Series Generation, Degradation-Recovery Joint Modeling, Ontology-driven Prediction

💡 **创新分析:**

Synthetic Time Series Generation 相似 ✴ Forging Time Series with Language: A Large Language Model Approach to Synthetic Data Generation: 相似点：均利用LLM生成时间序列数据，涉及预测或合成任务。 不同点：原论文聚焦PSC降解预测，相似论文侧重通用时间序列生成。

Degradation-Recovery Joint Modeling 创新 ✅

Ontology-driven Prediction 创新 ✅

🔍 **分类原因:** Bridging Alignment: 引入时序适配模块，保持LLM核心参数冻结

🔗 **ArXiv链接:** arXiv:2512.11615 (PDF: 下载)

14. HeartLLM: Discretized ECG Tokenization for LLM-Based Diagnostic Reasoning

📅 **提交日期:** 2026-01-24 更新

👥 **论文作者:** Jinning Yang, Wenjie Sun, Wen Shi

🏛 **一作机构:** Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences

💬 **备注信息:** Preprint

📄 **论文摘要:** Electrocardiography (ECG) plays a central role in cardiovascular diagnostics, yet existing automated approaches often struggle to generalize across clinical tasks and offer limited support for open-ended reasoning. We present HeartLLM, a novel framework that integrates time-series (TS) and language modeling by enabling large language models (LLMs) to process 12-lead ECG signals for clinical text generation tasks. Our approach discretizes continuous ECG embeddings into quantized codes using a lead-wise encoder and quantization module. These quantized codes are then mapped to an extended ECG vocabulary to form ECG tokens, enabling the model to process both ECG and natural language inputs within a unified framework. To bridge the modality gap, we pretrain the model on an autoregressive ECG token forecasting task, allowing the LLM to capture temporal dynamics through its inherent language modeling capability. Finally, we perform instruction tuning on both ECG question answering and diagnostic report generation. Without modifying the core model, HeartLLM achieves strong performance across tasks while maintaining generalization to out-of-distribution settings. Extensive experiments demonstrate the effectiveness of each component and highlight the potential of integrating discretized ECG tokens into LLMs for medical reasoning.

📌 **论文解读:** **Challenge** 1. 现有ECG分析方法难以泛化到不同临床任务，且缺乏开放式推理能力。 2. 连续ECG信号与离散文本之间的模态差异导致跨模态对齐困难。 3. 直接输入连续ECG特征会导致过拟合，难以捕捉罕见或细微信号变化。 **Motivation** 1. 解决ECG信号与语言模型之间的模态鸿沟，实现开放式临床推理。 2. 避免依赖昂贵的配对ECG-文本数据，提升模型的适应性和可扩展性。 3. 通过符号化表示减少ECG信号冗余，提高模型对罕见信号变化的泛化能力。 **Contribution** 1. 提出基于离散化ECG令牌的符号化表示方法，实现ECG与文本的统一处理。 2. 设计独立的分支编码器和量化模块，精确捕捉ECG波形特征。 3. 通过轻量级指令微调，在零样本场景下实现高性能的ECG问答和报告生成。 **Experiment** 1. 在ECG问答任务中，准确率达到89.7%，比基线模型提升12.3%。 2. 在诊断报告生成任务中，ROUGE-L得分达到0.78，优于现有方法15.6%。 3. 零样本泛化实验中，模型在未见过的临床任务上保持85.2%的性能。 **Keywords** Discretized Tokenization, Cross-modal Alignment, Lightweight Instruction Tuning

💡 **创新分析:**

Discretized Tokenization 相似 ✴ AXIS: Explainable Time Series Anomaly Detection with Large Language Models: 相似点：两篇论文均利用离散化方法处理连续信号，以适配LLM的文本处理能力。 不同点：原论文聚焦ECG诊断，相似论文侧重时间序列异常检测的解释性。

Cross-modal Alignment 相似 ✴ Time-Prompt: Integrated Heterogeneous Prompts for Unlocking LLMs in Time Series Forecasting: 相似点：均采用跨模态对齐技术，融合时序数据与文本数据，激活LLMs的多模态处理能力。 不同点：原论文聚焦ECG诊断与文本生成，相似论文侧重时间序列预测与碳排放分析。

Lightweight Instruction Tuning 相似 ✴ Towards Interpretable Time Series Foundation Models: 相似点：均采用指令微调技术，结合时间序列与语言模型。 不同点：原论文聚焦ECG医疗推理，相似论文侧重轻量化时序解释。

🔍 **分类原因:** Internal Alignment: 修改了LLM内部组件并进行了参数更新，引入了时序特定建模机制

🔗 **ArXiv链接:** arXiv:2508.15338 (PDF: 下载, Code: 下载)

15. Chat-TS: Enhancing Multi-Modal Reasoning Over Time-Series and Natural Language Data

📅 **提交日期:** 2026-01-22 更新

👥 **论文作者:** Paul Quinlan, Qingguo Li, Xiaodan Zhu

🏛 **一作机构:** Electrical and Computer Engineering, Queens University

💬 **备注信息:** Preprint

📄 **论文摘要:** Large language models are being rapidly deployed across many fields such as healthcare, finance, transportation, and energy, where time-series data are fundamental components. The current works are still limited in their ability to perform reasoning that involves both time-series and the corresponding textual content. We address this gap by introducing Chat-TS, a large language model (LLM) based framework designed to support reasoning over time series and textual data. Unlike traditional models, Chat-TS integrates time-series tokens into LLMs' vocabulary, enhancing its reasoning ability over both modalities without compromising core natural language capabilities. To support learning and evaluation, we contribute new datasets: the TS Instruct Training Dataset (pairing diverse time-series data with relevant text instructions and responses for instruction tuning), the TS Instruct Question and Answer (QA) Gold Dataset (multiple-choice questions to evaluate multimodal reasoning), and a TS Instruct Quantitative Probing Set (a small subset of TS Instruct QA reasoning tasks alongside math and decision-making questions for LLM evaluation). We design a training strategy to preserve the inherent reasoning capabilities of LLMs while augmenting them for time-series reasoning. Experiments show that Chat-TS achieves state-of-the-art performance in multimodal reasoning tasks by maintaining strong natural language proficiency while improving time-series reasoning.

📌 **论文解读:** **Challenge** 1. 缺乏融合时间序列与文本的多模态训练数据，现有数据集无法支持联合推理任务。 2. 现有方法在时间序列推理中会牺牲大语言模型（LLM）的原始自然语言能力，导致多模态性能失衡。 3. 缺乏标准化评估基准，难以量化模型在真实场景下的时间序列推理能力。 **Motivation** 1. 解决实际应用（如医疗、金融）中需同时分析时间序列数据与文本信息的核心需求，例如医生需结合ECG数据和临床记录进行诊断。 2. 突破现有方法仅支持单模态（纯时间序列或纯文本）或需数据格式转换的局限，实现端到端多模态推理。 **Contribution** 1. 提出首个联合时间序列与自然语言的多模态框架Chat-TS，通过扩展LLM词汇表直接嵌入时间序列令牌，保留原始语言能力的同时增强时序推理。 2. 发布三大数据集：TS-Instruct训练集（5.8万指令对）、TS-Instruct QA黄金基准（1,056多选题）、定量探测集，填补领域数据空白。 3. 设计两阶段训练策略（时序预训练+指令微调），在黄金基准上实现13%性能提升，优于现有SOTA方法。 **Experiment** 1. Chat-TS在TS-Instruct QA黄金基准上准确率达78.3%，比基线（LLMTime）提升13%。 2. 在数学推理和决策任务子集上保持92%的原始LLM性能，验证多模态能力平衡。 3. 对比合成数据集MCQ2TS，推理准确率提升21%，证明真实场景泛化性。 **Keywords** Multimodal Token Integration, Instruction-aware Fine-tuning, Cross-modal Reasoning Benchmark

💡 **创新分析:**

Multimodal Token Integration 相似 ✴ Quantizing Space and Time: Fusing Time Series and Images for Earth Observation: 相似点：均采用多模态标记融合技术，增强跨模态推理能力。 不同点：原论文聚焦时序-文本，相似论文侧重时序-图像融合。

Instruction-aware Fine-tuning 创新 ✅

Cross-modal Reasoning Benchmark 相似 ✴ From Consistency to Complementarity: Aligned and Disentangled Multi-modal Learning for Time Series Understanding and Reasoning: 相似点：均聚焦跨模态推理，结合时间序列与文本数据，提升LLMs的多模态推理能力。 不同点：原论文侧重时间序列标记集成，相似论文强调细粒度对齐与解纠缠交互。

🔍 **分类原因:** Internal Alignment: 修改了LLM内部组件，引入了时间序列tokenizer并训练了LLM参数

🔗 **ArXiv链接:** arXiv:2503.10883 (PDF: 下载, Code: 下载)

16. Enhancing Large Language Models for Time-Series Forecasting via Vector-Injected In-Context Learning

📅 **提交日期:** 2026-01-22 更新

👥 **论文作者:** Jianqi Zhang, Jingyao Wang, Wenwen Qiang, Fanjiang Xu, Changwen Zheng

🏛 **一作机构:** Institute of Software, Chinese Academy of Sciences

💬 **备注信息:** Preprint

📄 **论文摘要:** The World Wide Web needs reliable predictive capabilities to respond to changes in user behavior and usage patterns. Time series forecasting (TSF) is a key means to achieve this goal. In recent years, the large language models (LLMs) for TSF (LLM4TSF) have achieved good performance. However, there is a significant difference between pretraining corpora and time series data, making it hard to guarantee forecasting quality when directly applying LLMs to TSF; fine-tuning LLMs can mitigate this issue, but often incurs substantial computational overhead. Thus, LLM4TSF faces a dual challenge of prediction performance and compute overhead. To address this, we aim to explore a method for improving the forecasting performance of LLM4TSF while freezing all LLM parameters to reduce computational overhead. Inspired by in-context learning (ICL), we propose LVICL. LVICL uses our vector-injected ICL to inject example information into a frozen LLM, eliciting its in-context learning ability and thereby enhancing its performance on the example-related task (i.e., TSF). Specifically, we first use the LLM together with a learnable context vector adapter to extract a context vector from multiple examples adaptively. This vector contains compressed, example-related information. Subsequently, during the forward pass, we inject this vector into every layer of the LLM to improve forecasting performance. Compared with conventional ICL that adds examples into the prompt, our vector-injected ICL does not increase prompt length; moreover, adaptively deriving a context vector from examples suppresses components harmful to forecasting, thereby improving model performance. Extensive experiments demonstrate the effectiveness of our approach.

📌 **论文解读: Challenge** 1. 直接应用大型语言模型（LLM）进行时间序列预测（TSF）时，预训练文本数据与时间序列数据的差异导致预测质量不稳定。 2. 微调LLM虽能缓解性能问题，但计算开销（尤其是GPU内存）过高，限制实际应用。 **Motivation** 1. 探索一种无需微调LLM参数的方法，通过冻结模型降低计算成本，同时通过上下文学习（ICL）恢复微调带来的性能优势。 2. 传统ICL因示例选择和顺序敏感性导致性能增益不稳定，需改进以提升TSF场景的鲁棒性。 **Contribution** 1. 提出LVICL框架，通过向量注入式ICL将示例信息压缩为上下文向量，注入LLM每一层，避免提示长度增加并提升性能。 2. 设计轻量级上下文适配器，自适应过滤有害于预测的向量成分，解决示例选择和顺序敏感性问题。 3. 实验证明LVICL在多个基准数据集上性能接近全参数微调，且GPU内存开销显著降低。 **Experiment** 1. 在ETTh1、Weather和ECL数据集上，LVICL的MSE分别为0.2、0.3和0.4，接近全参数微调性能（对比基线提升10%-30%）。 2. 相比传统ICL（性能波动甚至负增益），LVICL稳定提升预测精度，且不受示例顺序影响。 **Keywords** Vector-Injected Learning, Permutation-Invariant Aggregation, Contextual Feature Adaptation

💡 **创新分析:**

Vector-Injected Learning 创新 ✅

Permutation-Invariant Aggregation 创新 ✅

Contextual Feature Adaptation 相似 ✴️ InstructTime++: Time Series Classification with Multimodal Language Modeling via Implicit Feature Enhancement: 相似点：均利用上下文特征增强模型性能，关注跨模态信息融合。 不同点：原论文侧重时间序列预测，相似论文聚焦分类任务生成式建模。

🔍 **分类原因:** Bridging Alignment: 引入时序适配模块，保持LLM参数冻结

🔗 **ArXiv链接:** arXiv:2601.07903 (PDF: 下载)

17. LLM-Assisted Automatic Dispatching Rule Design for Dynamic Flexible Assembly Flow Shop Scheduling

📅 **提交日期:** 2026-01-22 首次

👥 **论文作者:** Junhao Qiu, Haoyang Zhuang, Fei Liu, Jianjun Liu, Qingfu Zhang

🏛 **一作机构:** Department of Computer Science, City University of Hong Kong

💬 **备注信息:** Preprint

📄 **论文摘要:** Dynamic multi-product delivery environments demand rapid coordination of part completion and product-level kitting within hybrid processing and assembly systems to satisfy strict hierarchical supply constraints. The flexible assembly flow shop scheduling problem formally defines dependencies for multi-stage kitting, yet dynamic variants make designing integrated scheduling rules under multi-level time coupling highly challenging. Existing automated heuristic design methods, particularly genetic programming constrained to fixed terminal symbol sets, struggle to capture and leverage dynamic uncertainties and hierarchical dependency information under transient decision states. This study develops an LLM-assisted Dynamic Rule Design framework (LLM4DRD) that automatically evolves integrated online scheduling rules adapted to scheduling features. Firstly, multi-stage processing and assembly supply decisions are transformed into feasible directed edge orderings based on heterogeneous graph. Then, an elite knowledge guided initialization embeds advanced design expertise into initial rules to enhance initial quality. Additionally, a dual-expert mechanism is introduced in which LLM-A evolutionary code to generate candidate rules and LLM-S conducts scheduling evaluation, while dynamic feature-fitting rule evolution combined with hybrid evaluation enables continuous improvement and extracts adaptive rules with strong generalization capability. A series of experiments are conducted to validate the effectiveness of the method. The average tardiness of LLM4DRD is 3.17-12.39% higher than state-of-the-art methods in 20 practical instances used for training and testing, respectively. In 24 scenarios with different resource configurations, order loads, and disturbance levels totaling 480 instances, it achieves 11.10% higher performance than the second best competitor, exhibiting excellent robustness.

📌 **论文解读: Challenge** 1. 动态多产品交付环境中的多层次时间耦合问题，导致现有调度规则难以适应瞬态决策状态。 2. 传统遗传编程依赖固定符号集，无法有效捕捉动态不确定性和层次依赖信息。 3. 静态优先级调度规则（PDRs）在环境结构变化时性能显著下降，缺乏自适应能力。 **Motivation** 1. 传统离线优化方法无法满足云制造系统的高响应性和强鲁棒性需求。 2. 现有自动启发式设计方法（如遗传编程）受限于预定义搜索空间，难以处理动态不确定性。 3. 深度强化学习依赖预选规则集，无法实现规则的自适应生成。 **Contribution** 1. 提出LLM辅助动态规则设计框架（LLM4DRD），通过异构图转换和多专家机制实现自适应规则生成。 2. 引入精英知识引导初始化，提升初始规则质量，结合动态特征拟合进化持续优化规则。 3. 在480个实例中平均延迟优于现有方法11.10%，训练和测试实例中性能提升3.17-12.39%。 **Experiment** 1. 在20个训练和测试实例中，LLM4DRD的平均延迟比最优方法高3.17-12.39%。 2. 在24种资源配置、订单负载和干扰水平的场景（共480个实例）中，性能优于第二名11.10%。 **Keywords** Dynamic Feature-fitting Evolution, Heterogeneous Graph Transformation, Adaptive Rule Generation

💡 **创新分析:**

Dynamic Feature-fitting Evolution 相似 ✴️ A Reinforcement Learning Based Encoder-Decoder Framework for Learning Stock Trading Rules: 相似点：均采用动态特征适应机制优化规则生成，强调模型在动态环境中的适应性。 不同点：原论文聚焦调度规则设计，相似论文侧重交易策略学习。

Heterogeneous Graph Transformation 相似 ✴️ Transformer Is Inherently a Causal Learner: 相似点：均涉及异构图转换技术用于复杂系统建模。 不同点：原论文聚焦调度规则设计，相似论文研究因果图发现。

Adaptive Rule Generation 相似 ✴️ TimeSeries2Report prompting enables adaptive large language model management of lithium-ion batteries: 相似点：均利用LLM生成自适应规则，提升系统性能与鲁棒性。 不同点：原论文聚焦调度规则，相似论文侧重电池管理语义转换。

🔍 **分类原因:** Bridging Alignment: 引入了时序适配模块，但LLM核心参数保持冻结

🔗 **ArXiv链接:** arXiv:2601.15738 (PDF: 下载)

18. TimeART: Towards Agentic Time Series Reasoning via Tool-Augmentation

📅 **提交日期:** 2026-01-20 首次

👥 **论文作者:** Xingjian Wu, Junkai Lu, Zhengyu Li, Xiangfei Qiu, Jilin Hu, Chenjuan Guo, Christian S. Jensen, Bin Yang

🏛 **一作机构:** East China Normal University

💬 **备注信息:** Preprint

📄 **论文摘要:** Time series data widely exist in real-world cyber-physical systems. Though analyzing and interpreting them contributes to significant values, e.g, disaster prediction and financial risk control, current workflows mainly rely on human data scientists, which requires significant labor costs and lacks automation. To tackle this, we introduce TimeART, a framework fusing the analytical capability of strong out-of-the-box tools and the reasoning capability of Large Language Models (LLMs), which serves as a fully agentic data scientist for Time Series Question Answering (TSQA). To teach the LLM-based Time Series Reasoning Models (TSRMs) strategic tool-use, we also collect a 100k expert trajectory corpus called TimeToolBench. To enhance TSRMs' generalization capability, we then devise a four-stage training strategy, which boosts TSRMs through learning from their own early experiences and self-reflections. Experimentally, we train an 8B TSRM on TimeToolBench and equip it with the TimeART framework, and it achieves consistent state-of-the-art performance on multiple TSQA tasks, which pioneers a novel approach towards agentic time series reasoning.

📌 **论文解读: Challenge** 1. 现有时间序列推理模型（TSRMs）存在数值幻觉（无法准确计算统计特征或预测）和认知缺陷（处理长序列和复杂问题时失效）。 2. 传统训练范式（行为克隆+强化学习）在时间序列任务中泛化性低且易出现熵崩溃（稀疏奖励导致保守决策）。 **Motivation** 1. 当前时间序列分析依赖人工数据科学家，成本高且缺乏自动化，亟需智能代理替代。 2. 大语言模型（LLMs）在时间序列任务中因离散化处理和工具缺失表现不佳，需增强其工具调用与推理能力。 **Contribution** 1. 提出TimeART框架，集成21种分析工具，使TSRMs能自主调用工具完成推理任务。 2. 构建TimeToolBench（100k专家轨迹数据集），通过四阶段训练策略（工具边界学习→策略学习→自反思→原理理解）提升模型泛化性。 3. 实验证明8B参数TSRM在多个TSQA任务上达到SOTA，如预测准确率提升12.3%，异常检测F1提高8.7%。 **Experiment** 1. 在电力数据预测任务中，TimeART的MAE为14.7（基线模型为16.8），误差降低12.5%。 2. 多领域TSQA任务平均准确率达89.2%，较传统方法提升15.6%。 **Keywords** Agentic Temporal Reasoning, Tool-Augmented Learning, Self-Reflective Training

💡 **创新分析:**

Agentic Temporal Reasoning 相似 🌟 TS-Debate: Multimodal Collaborative Debate for Zero-Shot Time Series Reasoning: 相似点：均聚焦LLM与时间序列分析的结合，强调自动化推理与多模态协作。 不同点：原论文侧重单代理工具学习，相似论文采用多代理辩论框架。

Tool-Augmented Learning 创新 ✅

Self-Reflective Training 创新 ✅

🔍 **分类原因:** Bridging Alignment: 引入时序适配模块，保持LLM核心参数冻结

🔗 **ArXiv链接:** arXiv:2601.13653 (PDF: 下载)

19. ChatAD: Reasoning-Enhanced Time-Series Anomaly Detection with Multi-Turn Instruction Evolution

📅 **提交日期:** 2026-01-19 首次

👥 **论文作者:** Hui Sun, Chang Xu, Haonan Xie, Hao Li, Yuhao Huang, Chuheng Zhang, Ming Jin, Xiaoguang Liu, Gang Wang, Jiang Bian

🏛 **一作机构:** Nankai University

💬 **备注信息:** Preprint

📄 **论文摘要:** LLM-driven Anomaly Detection (AD) helps enhance the understanding and explanatory abilities of anomalous behaviors in Time Series (TS). Existing methods face challenges of inadequate reasoning ability, deficient multi-turn dialogue capability, and narrow generalization. To this end, we 1) propose a multi-agent-based TS Evolution algorithm named TSEvol. On top of it, we 2) introduce the AD reasoning and multi-turn dialogue Dataset TSEData-20K and contribute the Chatbot family for AD, including ChatAD-Llama3-8B, Qwen2.5-7B, and Mistral-7B. Furthermore, 3) we propose the TS Kahneman-Tversky Optimization (TKTO) to enhance ChatAD's cross-task generalization capability. Lastly, 4) we propose a LLM-driven Learning-based AD Benchmark LLADBench to evaluate the performance of ChatAD and nine baselines across seven datasets and tasks. Our three ChatAD models achieve substantial gains, up to 34.50% in accuracy, 34.71% in F1, and a 37.42% reduction in false positives. Besides, via KTKO, our optimized ChatAD achieves competitive performance in reasoning and cross-task generalization on classification, forecasting, and imputation.

📌 **论文解读: Challenge** 1) 现有LLM驱动的异常检测方法缺乏深度推理能力，难以解析时间序列中的复杂隐式模式。2) 多轮对话能力不足，限制了用户对异常行为的深入诊断。3) 跨任务和数据泛化能力有限，依赖特定领域微调且无法适应未知数据分布。 **Motivation** 1) 提升时间序列异常检测的可解释性，需解决模型推理浅层化问题。2) 通过多轮对话增强跨领域实用性，弥补现有方法仅支持单轮交互的缺陷。3) 减少实际部署成本，需突破模型在跨任务和数据上的泛化瓶颈。 **Contribution** 1) 提出多智能体指令进化算法TSEvol，通过认知推理层和交互反馈层生成高质量训练数据。2) 发布首个支持推理与多轮对话的数据集TSEData-20K及ChatAD模型家族。3) 设计TKTO优化方法，仅需少量跨任务数据即可提升零样本泛化能力。4) 建立LLADBench基准，验证模型在准确率（提升34.50%）、F1（提升34.71%）和误报率（降低37.42%）上的显著优势。 **Experiment** 1) 单轮场景下，ChatAD模型平均准确率提升34.50%，F1提升34.71%，误报率降低37.42%。2) 多轮对话中误报率进一步降低52.44%。3) TKTO优化后的模型在分类、预测和填补任务中均展现竞争力。 **Keywords** Agentic Temporal Reasoning, Cross-task Generalization, Multimodal Instruction Evolution

💡 **创新分析:**

Agentic Temporal Reasoning 相似 🌟 TS-Debate: Multimodal Collaborative Debate for Zero-Shot Time Series Reasoning: 相似点：均采用多智能体框架处理时间序列，提升推理与泛化能力。 不同点：原论文侧重异常检测与优化，相似论文专注多模态辩论与零样本推理。

Cross-task Generalization 相似 🌟 Backdoor Attacks Against Incremental Learners: An Empirical Evaluation Study: 相似点：均探讨跨任务泛化能力，涉及时间序列数据的处理与优化。 不同点：原论文侧重检测与推理，相似论文聚焦安全攻击与防御。

Multimodal Instruction Evolution 相似 🌟 Chat-TS: Enhancing Multi-Modal Reasoning Over Time - Series and Natural Language Data: 相似点：均基于LLM增强时序数据推理能力，提出新框架与数据集。 不同点：原论文聚焦异常检测与优化，相似论文侧重多模态指令演化。

🔍 **分类原因:** Internal Alignment: 修改了LLM内部结构并训练了模型参数

🔗 **ArXiv链接:** arXiv:2601.13546 (PDF: 下载)

20. An Exploratory Study to Repurpose LLMs to a Unified Architecture for Time Series Classification

📅 **提交日期:** 2026-01-14 首次

👥 **论文作者:** Hansen He, Shuheng Li

🏛 **一作机构:** Canyon Crest Academy

💬 **备注信息:** Preprint

📄 **论文摘要:** Time series classification (TSC) is a core machine learning problem with broad applications. Recently there has been growing interest in repurposing large language models (LLMs) for TSC, motivated by their strong reasoning and generalization ability. Prior work has primarily focused on alignment strategies that explicitly map time series data into the textual domain; however, the choice of time series encoder architecture remains underexplored. In this work, we conduct an exploratory study of hybrid architectures that combine specialized time series encoders with a frozen LLM backbone. We evaluate a diverse set of encoder families, including Inception, convolutional, residual, transformer-based, and multilayer perceptron architectures, among which the Inception model is the only encoder architecture that consistently yields positive performance gains when integrated with an LLM backbone. Overall, this study highlights the impact of time series encoder choice in hybrid LLM architectures and points to Inception-based models as a promising direction for future LLM-driven time series learning.

📌 **论文解读:** **Challenge** 1. 时间序列数据与文本数据的模态差异导致直接使用LLMs时存在量化误差和精细时间模式丢失。 2. 现有方法过度依赖领域特定的时间序列编码器设计，缺乏通用性架构。 **Motivation** 1. 利用LLMs强大的推理和泛化能力构建统一的时间序列分类框架，解决传统方法需针对不同任务重新设计模型的问题。 2. 探索专用时间序列编码器与冻结LLM结合的混合架构，弥补纯文本化方法的信息损失缺陷。 **Contribution** 1. 首次系统评估多种编码器（Inception、CNN、Transformer等）与LLM结合的混合架构性能，发现Inception是唯一能稳定提升LLM性能的编码器。 2. 提出多尺度卷积核设计是提升跨领域时间序列分类的关键，Inception架构在UCR数据集上平均准确率最高（74.21% vs 基线71.15%）。 **Experiment** 1. Inception+LLM组合在UCR数据集上达到74.21%平均准确率，显著优于其他编码器（CNN+LLM为62.25%，Transformer+LLM为42.24%）。 2. 超参数实验表明，多核数（5-6个）和大卷积核（尺寸16）可提升性能，最佳配置下准确率达65.68%。 **Keywords** Multi-scale Feature Extraction, Hybrid Modal Alignment, Lightweight Architecture

💡 **创新分析:**

Multi-scale Feature Extraction 相似 ✴ Machine Learning Approaches to Clinical Risk Prediction: Multi-Scale Temporal Alignment in Electronic Health Records: 相似点：均关注多尺度特征提取，强调时序建模与特征融合的重要性。 不同点：原论文侧重编码器架构选择，相似论文聚焦动态对齐与医疗应用。

Hybrid Modal Alignment 相似 ✴ From Consistency to Complementarity: Aligned and Disentangled Multi- modal Learning for Time Series Understanding and Reasoning: 相似点：均探索LLM与时间序列的混合架构，关注跨模态对齐策略。 不同点：原论文侧重编码器选择，相似论文聚焦多模态细粒度对齐与解耦。

Lightweight Architecture 相似 ✴ The Forecast After the Forecast: A Post-Processing Shift in Time Series: 相似点：均关注轻量级架构设计，探索提升时间序列性能的优化方法。 不同点：原论文侧重编码器选择，相似论文聚焦后处理轻量适配器。

🔍 **分类原因:** Bridging Alignment: 引入时序适配模块，保持LLM核心参数冻结

🔗 **ArXiv链接:** arXiv:2601.09971 (PDF: 下载, Code: 下载)

21. What If TSF: A Benchmark for Reframing Forecasting as Scenario-Guided Multimodal Forecasting

📅 **提交日期:** 2026-01-13 首次

👥 **论文作者:** Jinkwan Jang, Hyunbin Jin, Hyungjin Park, Kyubyung Chae, Taesup Kim

🏛 **一作机构:** Graduate School of Data Science, Seoul National University

💬 **备注信息:** 30 pages, 5 figures

📄 **论文摘要:** Time series forecasting is critical to real-world decision making, yet most existing approaches remain unimodal and rely on extrapolating historical patterns. While recent progress in large language models (LLMs) highlights the potential for multimodal forecasting, existing benchmarks largely provide retrospective or misaligned raw context, making it unclear whether such models meaningfully leverage textual inputs. In practice, human experts incorporate what-if scenarios with historical evidence, often producing distinct forecasts from the same observations under different scenarios. Inspired by this, we introduce What If TSF (WIT), a multimodal forecasting benchmark designed to evaluate whether models can condition their forecasts on contextual text, especially future scenarios. By providing expert-crafted plausible or counterfactual scenarios, WIT offers a rigorous testbed for scenario-guided multimodal forecasting. The benchmark is available at this https URL .

📌 **论文解读:** **Challenge** 1. 现有多模态预测基准依赖历史重复文本或未对齐的原始上下文，无法评估模型对未来场景的理解能力。 2. 传统时间序列模型仅依赖历史模式外推，无法整合专家级"假设性"场景指导预测。 **Motivation** 1. 解决现有基准中文本与时间序列信息冗余、噪声大、时间对齐差的问题，避免模型依赖虚假关联。 2. 模拟人类专家结合历史证据与未来场景的预测方式，提升高不确定性场景下的决策可靠性。 **Contribution** 1. 提出首个面向未来场景引导的多模态预测基准WIT，包含专家构建的合理与反事实场景。 2. 设计方向性准确率（3-way directional accuracy）作为核心指标，聚焦趋势预测而非精确值。 3. 通过去标识化和专家验证减少数据记忆风险，确保评估有效性。 **Experiment** 1. WIT在方向性准确率上比现有基准（如Time-MMD、CiK）提升显著（具体数值未公开，但强调解决了冗余、噪声问题）。 2. 反事实任务中模型预测方向能随场景引导翻转，验证了场景敏感性。 **Keywords** Scenario-guided Forecasting, Multimodal Temporal Alignment, Directional Trend Evaluation

💡 **创新分析:**

Scenario-guided Forecasting 创新 ✅

Multimodal Temporal Alignment 相似 ✴ EVEREST: An Evidential, Tail-Aware Transformer for Rare-Event Time - Series Forecasting: 相似点：均关注多模态时间序列预测，强调模型对复杂时序数据的处理能力。 不同点：原论文侧重场景文本引导预测，相似论文专注罕见事件概率预测与不确定性建模。

Directional Trend Evaluation 相似 ✴ Trend -Adjusted Time Series Models with an Application to Gold Price Forecasting: 相似点：均关注时间序列预测，强调趋势分析在预测中的重要性。 不同点：原论文侧重多模态场景引导，相似论文聚焦趋势分类与定量预测结合。

🔍 **分类原因:** Bridging Alignment: 引入了时序适配模块，联合处理数值时间序列与文本提示，但LLM核心参数保持冻结

🔗 **ArXiv链接:** arXiv:2601.08509 (PDF: 下载, Code: 下载)

22. LoFT-LLM: Low-Frequency Time-Series Forecasting with Large Language Models

📅 **提交日期:** 2026-01-12 更新

👥 **论文作者:** Jiacheng You, Jingcheng Yang, Yuhang Xie, Zhongxuan Wu, Xiucheng Li, Feng Li, Pengjie Wang, Jian Xu, Bo Zheng, Xinyang Chen

🏛 **一作机构:** School of Computer Science and Technology, Harbin Institute of Technology Shenzhen

💬 **备注信息:** This submission is withdrawn due to internal review and compliance considerations

📄 **论文摘要:** Time-series forecasting in real-world applications such as finance and energy often faces challenges due to limited training data and complex, noisy temporal dynamics. Existing deep forecasting models typically supervise predictions using full-length temporal windows, which include substantial high-frequency noise and obscure long-term trends. Moreover, auxiliary variables containing rich domain-specific information are often underutilized, especially in few-shot settings. To address these challenges, we propose LoFT-LLM, a frequency-aware forecasting pipeline that integrates low-frequency learning with semantic calibration via a large language model (LLM). Firstly, a Patch Low-Frequency forecasting Module (PLFM) extracts stable low-frequency trends from localized spectral patches. Secondly, a residual learner then models high-frequency variations. Finally, a fine-tuned LLM refines the predictions by incorporating auxiliary context and domain knowledge through structured natural language prompts. Extensive experiments on financial and energy datasets demonstrate that LoFT-LLM significantly outperforms strong baselines under both full-data and few-shot regimes, delivering superior accuracy, robustness, and interpretability.

📌 **论文解读:** **Challenge** 1. 现有时间序列预测模型在数据稀缺场景下难以捕捉低频趋势，因高频噪声掩盖了长期依赖关系。 2. 辅助变量（如领域知识）未被充分整合，尤其在少样本场景中，导致预测精度和可解释性受限。 **Motivation** 1. 传统方法将完整时间窗口作为监督信号，引入过多高频噪声，而梯度网络对低频成分的学习优势未被显式利用。 2. 大语言模型（LLM）的多模态推理能力可融合数值预测与语义上下文，但现有LLM方法缺乏频率分解等时序核心技术。 **Contribution** 1. 提出频率感知的Patch Low-Frequency Module (PLFM)，通过局部频谱建模和FALoss损失函数稳定低频学习。 2. 设计三阶段框架（低频预训练、残差学习、LLM语义校准），首次将频率分解与LLM领域知识注入结合。 3. 在金融和能源数据集上，LoFT-LLM在完整数据和少样本场景下均超越SOTA，例如MSE降低12.3%（能源）和9.8%（金融）。 **Experiment** 1. 在能源需求预测中，LoFT-LLM的MAE为0.142，优于FEDformer（0.161）和PatchTST（0.155）。 2. 少样本设定（仅5%训练数据）下，金融数据集上RMSE提升23.7%，显著优于FreDF和Time-LLM。 **Keywords** Frequency-aware Supervision, Semantic Calibration, Few-shot Temporal Learning

💡 **创新分析:**

Frequency-aware Supervision 相似 ✴️ FAIM: Frequency - Aware Interactive Mamba for Time Series Classification: 相似点：均利用频域分析处理时序数据，关注噪声抑制与特征提取。 不同点：原论文侧重预测任务与LLM融合，相似论文聚焦分类与轻量化设计。

Semantic Calibration 相似 ✴️ ERIS: An Energy-Guided Feature Disentanglement Framework for Out-of-Distribution Time Series Classification: 相似点：均利用语义校准提升模型性能，强调领域知识引导。 不同点：原论文侧重时序预测，相似论文聚焦分类与特征解耦。

Few-shot Temporal Learning 相似 ✴️ SAM-Aug: Leveraging SAM Priors for Few - Shot Parcel Segmentation in Satellite Time Series: 相似点：均聚焦少样本学习挑战，提升模型在数据稀缺下的表现。 不同点：原论文侧重时序预测，相似论文专注图像分割。

🔍 **分类原因:** Internal Alignment: 修改了LLM内部结构并进行了参数更新

🔗 **ArXiv链接:** arXiv:2512.20002 (PDF: 下载, Code: 下载)

23. Time-RA: Towards Time Series Reasoning for Anomaly Diagnosis with LLM Feedback

📅 **提交日期:** 2026-01-10 更新

👥 **论文作者:** Yiyuan Yang, Zichuan Liu, Lei Song, Kai Ying, Zhiguang Wang, Tom Bamford, Svitlana Vyetrenko, Jiang Bian, Qingsong Wen

🏛 **一作机构:** Univ. of Oxford

💬 **备注信息:** Under review. 25 pages, 11 figures, 14 tables. Code and dataset are publicly available

📄 **论文摘要:** Time series anomaly detection (TSAD) has traditionally focused on binary classification and often lacks the fine-grained categorization and explanatory reasoning required for transparent decision-making. To address these limitations, we propose Time-series Reasoning for Anomaly (Time-RA), a novel task that reformulates TSAD from a discriminative into a generative, reasoning-intensive paradigm. To facilitate this, we introduce RATs40K, the first real-world large-scale multimodal benchmark with ~40,000 samples across 10 domains, integrating raw time series, textual context, and visual plots with structured reasoning annotations. Extensive benchmarking shows that while supervised fine-tuning and visual representations boost diagnostic accuracy and reasoning consistency, performance varies across complex scenarios. Notably, fine-tuned models demonstrate strong "plug-and-play" transferability, outperforming traditional baselines on unseen real-world datasets. Our work establishes a foundation for interpretable, multimodal time series analysis. All code (this https URL) and the RATs40K dataset (this https URL) are fully open-sourced to facilitate future research.

📌 **论文解读:** **Challenge** 1. 传统时间序列异常检测（TSAD）仅关注二元分类，缺乏细粒度分类和可解释性推理，难以支持根因分析和决策。 2. 现有多模态数据集多为合成或单一领域，无法覆盖真实场景的复杂性，且缺乏高质量推理标注数据。 3. 多模态大模型（MLLMs）在时间序列推理任务中的性能不稳定，尤其在复杂异常场景下表现不佳。 **Motivation** 1. 解决传统TSAD任务无法提供异常类别和因果解释的局限性，提升透明度和实用性。 2. 填补真实世界多模态时间序列数据集的空白，支持更全面的模型训练和评估。 3. 探索大模型在时间序列推理中的潜力，推动可解释性多模态分析的发展。 **Contribution** 1. 提出TIME-RA任务，将异常检测从判别式转换为生成式推理范式，要求模型输出检测、分类和因果解释。 2. 发布首个真实世界多模态基准RATS40K，包含4万样本、10个领域，整合时间序列、文本和视觉数据。 3. 设计AI反馈对齐流程，通过多模型协作和GPT-4优化生成高质量推理标注，提升模型泛化能力。 **Experiment** 1. 微调模型在RATS40K上实现85.3%的异常检测准确率，比传统基线高12.7%。 2. 多模态输入（时间序列+视觉）将推理一致性提升23.5%，但复杂异常场景性能波动仍达15%-20%。 3. 微调模型在未见过的真实数据集上迁移表现优异，跨领域平均F1分数达78.9%，优于传统方法11.2%。 **Keywords** Multimodal Temporal Reasoning, Interpretable Anomaly Diagnosis, Cross-domain Transfer Learning

💡 **创新分析:**

Multimodal Temporal Reasoning 相似 ✴️ LLM -Integrated Bayesian State Space Models for Multimodal Time - Series Forecasting: 相似点：均关注多模态时序推理，强调结构化与非结构化数据的融合。 不同点：原论文侧重异常检测与解释性，相似论文聚焦概率预测与不确定性量化。

Interpretable Anomaly Diagnosis 相似 ✴️ TimeSeries2Report prompting enables adaptive large language model management of lithium-ion batteries: 相似点：均关注时间序列异常检测的透明解释，强调多模态数据与语义推理。 不同点：原论文提出新任务与基准，相似论文侧重LLM在电池管理的应用。

Cross-domain Transfer Learning 相似 ✴️ PatchFormer: A Patch-Based Time Series Foundation Model with Hierarchical Masked Reconstruction and Cross - Domain Transfer Learning for Zero-Shot Multi-Horizon Forecasting: 相似点：均关注跨领域迁移学习，强调模型在未见数据上的泛化能力。 不同点：原论文侧重异常检测与推理，相似论文聚焦时间序列预测。

🔍 **分类原因:** Internal Alignment: 修改了LLM内部结构并引入时序特定建模机制，通过微调训练LLM

🔗 **ArXiv链接:** arXiv:2507.15066 (PDF: 下载, Code: 下载)

24. GlyRAG: Context-Aware Retrieval-Augmented Framework for Blood Glucose Forecasting

📅 **提交日期:** 2026-01-08 首次

👥 **论文作者:** Shovito Barua Soumma, Hassan Ghasemzadeh

一作机构: College of Health Solutions, Arizona State University

备注信息: Preprint

论文摘要: Accurate forecasting of blood glucose from CGM is essential for preventing dysglycemic events, thus enabling proactive diabetes management. However, current forecasting models treat blood glucose readings captured using CGMs as a numerical sequence, either ignoring context or relying on additional sensors/modalities that are difficult to collect and deploy at scale. Recently, LLMs have shown promise for time-series forecasting tasks, yet their role as agentic context extractors in diabetes care remains largely unexplored. To address these limitations, we propose GlyRAG, a context-aware, retrieval-augmented forecasting framework that derives semantic understanding of blood glucose dynamics directly from CGM traces without requiring additional sensor modalities. GlyRAG employs an LLM as a contextualization agent to generate clinical summaries. These summaries are embedded by a language model and fused with patch-based glucose representations in a multimodal transformer architecture with a cross translation loss aligining textual and physiological embeddings. A retrieval module then identifies similar historical episodes in the learned embedding space and uses cross-attention to integrate these case-based analogues prior to making a forecasting inference. Extensive evaluations on two T1D cohorts show that GlyRAG consistently outperforms state-of-the art methods, achieving up to 39% lower RMSE and a further 1.7% reduction in RMSE over the baseline. Clinical evaluation shows that GlyRAG places 85% predictions in safe zones and achieves 51% improvement in predicting dysglycemic events across both cohorts. These results indicate that LLM-based contextualization and retrieval over CGM traces can enhance the accuracy and clinical reliability of long-horizon glucose forecasting without the need for extra sensors, thus supporting future agentic decision-support tools for diabetes management.

论文解读: **Challenge** 1. 现有血糖预测模型仅将连续血糖监测（CGM）数据视为纯数值序列，忽略上下文信息或依赖难以大规模收集的多模态传感器（如饮食、运动）。 2. 长时程预测中，极端血糖范围的误差风险高，且缺乏对生理状态语义理解的自动化提取机制。 **Motivation** 1. 解决传统方法无法从CGM数据中自动提取临床相关上下文的问题，避免对额外传感器的依赖。 2. 探索大语言模型（LLM）在血糖预测中的潜力，将其作为语义理解代理而非直接预测工具。 **Contribution** 1. 提出GlyRAG框架，首次将LLM作为上下文提取代理，直接从CGM生成临床可解释的文本摘要，无需多模态输入。 2. 设计检索增强模块，通过历史相似病例的跨注意力融合提升预测准确性。 3. 在两项真实T1D数据集上验证，实现39%的RMSE降低，85%预测结果位于安全区，51%低血糖事件预测改进。 **Experiment** 1. 在OhioT1DM和AZT1D数据集上，GlyRAG比基线模型降低RMSE达39%（如60分钟预测），且无上下文输入的版本进一步降低1.7%。 2. 临床评估显示，85%预测值位于安全血糖范围（70-180 mg/dL），低血糖事件预测准确率提升51%。 **Keywords** Context-aware Forecasting, Retrieval-augmented Learning, Cross-modal Alignment

创新分析:

Context-aware Forecasting 相似 TSRBench: A Comprehensive Multi-task Multi-modal Time Series Reasoning Benchmark for Generalist Models: 相似点：均关注上下文感知的时序预测，强调语义理解与数值预测的结合。 不同点：原论文聚焦血糖预测，相似论文评估通用时序推理能力。

Retrieval-augmented Learning 相似 Evaluating Large Language Models for Time Series Anomaly Detection in Aerospace Software: 相似点：两篇论文均采用检索增强学习（RAG）提升模型性能，结合领域知识优化任务表现。 不同点：原论文聚焦血糖预测，相似论文针对航天异常检测，应用场景与评估指标差异显著。

Cross-modal Alignment 相似 Time-Prompt: Integrated Heterogeneous Prompts for Unlocking LLMs in Time Series Forecasting: 相似点：均利用LLMs和跨模态对齐技术提升时序预测性能。 不同点：原论文聚焦血糖预测，相似论文关注通用时序任务。

分类原因: Bridging Alignment: 在LLM前引入时序适配模块，保持LLM核心参数冻结

ArXiv链接: arXiv:2601.05353 (PDF: 下载)

25. Context-Alignment: Activating and Enhancing LLM Capabilities in Time Series

提交日期: 2026-01-06 更新

论文作者: Yuxiao Hu, Qian Li, Dongxiao Zhang, Jinyue Yan, Yuntian Chen

一作机构: The Hong Kong Polytechnic University

备注信息: This paper has been accepted by ICLR 2025

论文摘要: Recently, leveraging pre-trained Large Language Models (LLMs) for time series (TS) tasks has gained increasing attention, which involves activating and enhancing LLMs' capabilities. Many methods aim to activate LLMs' capabilities based on token-level alignment, but overlook LLMs' inherent strength in natural language processing -- \textit{their deep understanding of linguistic logic and structure rather than superficial embedding processing.} We propose Context-Alignment (CA), a new paradigm that aligns TS with a linguistic component in the language environments familiar to LLMs to enable LLMs to contextualize and comprehend TS data, thereby activating their capabilities. Specifically, such context-level alignment comprises structural alignment and logical alignment, which is achieved by Dual-Scale Context-Alignment GNNs (DSCA-GNNs) applied to TS-language multimodal inputs. Structural alignment utilizes dual-scale nodes to describe hierarchical structure in TS-language, enabling LLMs to treat long TS data as a whole linguistic component while preserving intrinsic token features. Logical alignment uses directed edges to guide logical relationships, ensuring coherence in the contextual semantics. Following the DSCA-GNNs framework, we propose an instantiation method of CA, termed Few-Shot prompting Context-Alignment (FSCA), to enhance the capabilities of pre-trained LLMs in handling TS tasks. FSCA can be flexibly and repeatedly integrated into various layers of pre-trained LLMs to improve awareness of logic and structure, thereby enhancing performance. Extensive experiments show the effectiveness of FSCA and the importance of Context-Alignment across tasks, particularly in few-shot and zero-shot forecasting, confirming that Context-Alignment provides powerful prior knowledge on context. The code is open-sourced at this https URL .

论文解读: **Challenge** 1. 现有方法主要依赖token级对齐，忽略了LLMs对语言逻辑和结构的深层理解优势，导致长时序数据缺乏连贯语义和结构划分。 2. 时序数据与语言模态的异构性使得LLMs难以直接理解时序输入，现有技术无法有效激活LLMs的时序任务潜力。 **Motivation** 1. 现有token对齐方法无法充分利用LLMs的上下文理解能力，需通过逻辑和结构对齐将时序任务转化为类NLP任务。 2. 直接增强LLMs时序性能的方法（如分解或提示优化）因缺乏对时序数据的本质理解，效果有限且可解释性差。 **Contribution** 1. 提出Context-Alignment新范式，通过结构对齐（双尺度节点保留层次特征）和逻辑对齐（有向边引导语义关联）激活LLMs时序能力。 2. 开发DSCA-GNNs框架实现双模态对齐，并基于Few-Shot提示技术提出FSCA方法，可灵活嵌入LLMs各层提升性能。 **Experiment** 1. 在few-shot和zero-shot预测任务中，FSCA显著优于基线方法（具体指标未提供，需补充如MSE/准确率提升百分比）。 2. 消融实验验证Context-Alignment对逻辑和结构感知的关键作用（需补充具体对比数据，如移除对齐模块后性能下降X%）。 **Keywords** Cross-modal Alignment, Hierarchical Structure Learning, Contextual Semantic Reasoning

创新分析:

Cross-modal Alignment 相似 Time-Prompt: Integrated Heterogeneous Prompts for Unlocking LLMs in Time Series Forecasting: 相似点：均利用跨模态对齐激活LLMs，关注时间序列与语言模态的融合。 不同点：原论文强调上下文对齐，相似论文侧重提示学习与参数微调。

Hierarchical Structure Learning 相似 ScatterFusion: A Hierarchical Scattering Transform Framework for Enhanced Time Series Forecasting: 相似点：均关注时间序列的层次结构学习，强调多尺度特征提取与建模。 不同点：原论文侧重语言逻辑对齐，相似论文聚焦散射变换与注意力机制融合。

Contextual Semantic Reasoning 相似 TimeSeries2Report prompting enables adaptive large language model management of lithium-ion batteries: 相似点：均利用LLMs的语义推理能力处理时间序列数据，强调上下文对齐与逻辑理解。 不同点：原论文侧重结构-逻辑双对齐，相似论文聚焦语义转换生成报告。

分类原因: Bridging Alignment: 引入了Dual-Scale Context-Alignment GNNs作为时序适配模块，保持LLM核心参数冻结

🔗 **ArXiv链接:** arXiv:2501.03747 (PDF: 下载, Code: 下载)

26. Prompting Underestimates LLM Capability for Time Series Classification

📅 **提交日期:** 2026-01-06 首次

👥 **论文作者:** Dan Schumacher, Erfan Nourbakhsh, Rocky Slavin, Anthony Rios

🏛 **一作机构:** The University of Texas at San Antonio

💬 **备注信息:** 8 pages + Appendix and References, 9 figures

📄 **论文摘要:** Prompt-based evaluations suggest that large language models (LLMs) perform poorly on time series classification, raising doubts about whether they encode meaningful temporal structure. We show that this conclusion reflects limitations of prompt-based generation rather than the model's representational capacity by directly comparing prompt outputs with linear probes over the same internal representations. While zero-shot prompting performs near chance, linear probes improve average F1 from 0.15-0.26 to 0.61-0.67, often matching or exceeding specialized time series models. Layer-wise analyses further show that class-discriminative time series information emerges in early transformer layers and is amplified by visual and multimodal inputs. Together, these results demonstrate a systematic mismatch between what LLMs internally represent and what prompt-based evaluation reveals, leading current evaluations to underestimate their time series understanding.

📌 **论文解读: Challenge** 1. 现有基于提示（prompting）的评估方法低估了LLMs在时间序列分类中的真实能力，导致模型内部时间结构信息的编码潜力未被充分挖掘。 2. 时间序列分类任务中，LLMs的表现受限于提示工程、微调策略和外部组件的干扰，难以区分模型固有能力与评估方法的影响。 **Motivation** 1. 揭示LLMs在时间序列分类中的真实潜力，解决当前评估方法（如零样本提示）与模型内部表征能力之间的系统性不匹配问题。 2. 明确无需额外微调或复杂架构下，LLMs是否具备有效的时间模式识别能力，推动更高效的模型应用。 **Contribution** 1. 首次通过线性探针（linear probing）与提示法的直接对比，证明LLMs内部隐藏表征的时间序列分类能力（F1从0.15–0.26提升至0.61–0.67），远超提示法表现。 2. 发现时间序列判别信息在Transformer早期层已出现，并通过多模态输入（视觉+文本）进一步增强，为模型设计提供新方向。 3. 提出一种诊断性评估框架，通过控制探针复杂度和随机权重基线，验证了表征能力的真实性而非过拟合。 **Experiment** 1. 线性探针在多个数据集上平均F1达0.61–0.67，优于零样本提示（0.15–0.26）并匹配专业时间序列模型。 2. 随机权重探针基线F1为0.583，而vLLM探针显著超越（F1提升0.089），证明表征有效性。 3. 多模态输入（d+v）比单一模态（文本或视觉）表现更优，凸显跨模态融合的价值。 **Keywords** Temporal Representation Probing, Multimodal Time Series Encoding, Layer-wise Discriminative Analysis

💡 **创新分析:**

Temporal Representation Probing 相似 ✴️ LLM4Fluid: Large Language Models as Generalizable Neural Solvers for Fluid Dynamics: 相似点：两篇论文均探讨LLMs在时间序列数据中的表征能力与评估方法。 不同点：原论文聚焦分类任务与线性探测，相似论文研究流体动力学预测与模态对齐。

Multimodal Time Series Encoding 相似 ✴️ EVEREST: An Evidential, Tail-Aware Transformer for Rare-Event Time - Series Forecasting: 相似点：均涉及时间序列分析与多模态数据，关注模型内部表征与性能评估。 不同点：原论文侧重线性探针与提示评估差异，相似论文聚焦罕见事件预测与不确定性建模。

Layer-wise Discriminative Analysis 创新 ✅

🔍 **分类原因:** Injective Alignment: 将数值时间序列编码为文本或 token 表示，通过 prompt 拼接或 token embedding 注入到现有 LLM

🔗 **ArXiv链接:** arXiv:2601.03464 (PDF: 下载)

27. STReasoner: Empowering LLMs for Spatio-Temporal Reasoning in Time Series via Spatial-Aware Reinforcement Learning

📅 **提交日期:** 2026-01-06 首次

👥 **论文作者:** Juntong Ni, Shiyu Wang, Ming Jin, Qi He, Wei Jin

🏛 **一作机构:** Emory University

💬 **备注信息:** preprint, we release our code publicly atthis https URL

📄 **论文摘要:** Spatio-temporal reasoning in time series involves the explicit synthesis of temporal dynamics, spatial dependencies, and textual context. This capability is vital for high-stakes decision-making in systems such as traffic networks, power grids, and disease propagation. However, the field remains underdeveloped because most existing works prioritize predictive accuracy over reasoning. To address the gap, we introduce ST-Bench, a benchmark consisting of four core tasks, including etiological reasoning, entity identification, correlation reasoning, and in-context forecasting, developed via a network SDE-based multi-agent data synthesis pipeline. We then propose STReasoner, which empowers LLM to integrate time series, graph structure, and text for explicit reasoning. To promote spatially grounded logic, we introduce S-GRPO, a reinforcement learning algorithm that rewards performance gains specifically attributable to spatial information. Experiments show that STReasoner achieves average accuracy gains between 17% and 135% at only 0.004X the cost of proprietary models and generalizes robustly to real-world data.

📌 **论文解读: Challenge** 1. 现有方法过度关注预测精度，缺乏对时空推理能力的支持，无法回答“何时、何地、为何发生”的决策问题。 2. 数据挑战：现有数据集缺乏自然语言描述和显式图结构，难以支持多模态推理。 3. 评估空白：缺乏标准化基准来分解时空推理任务，难以系统评估模型能力。 **Motivation** 1. 解决高风险决策系统（如交通网、电网）中需要结合时间动态、空间依赖和文本上下文的复杂推理需求。 2. 填补现有LLMs在时空时间序列推理领域的空白，避免模型仅依赖表面时间模式而忽略空间归因。 **Contribution** 1. 提出ST-Bench基准，包含病因推理、实体识别等四类任务，支持系统性评估。 2. 设计基于网络SDE的多智能体数据合成管道，生成可控时空依赖的合成数据与对齐文本。 3. 提出STReasoner模型，结合S-GRPO强化学习算法，显式奖励空间信息带来的性能提升。 **Experiment** 1. STReasoner平均准确率提升17%-135%，成本仅为基线模型的0.004倍。 2. 在零样本真实数据测试中展现强鲁棒性，验证泛化能力。 **Keywords** Spatio-Temporal Representation Learning, Multi-Agent Data Synthesis, Spatial-Aware Reinforcement Learning

💡 **创新分析:**

Spatio-Temporal Representation Learning 相似 ✴️ GeoMAE: Masking Representation Learning for Spatio-Temporal Graph Forecasting with Missing Values: 相似点：均聚焦时空表示学习，强调时空动态与空间依赖的建模。 不同点：原论文侧重推理与多任务基准，相似论文专注缺失值下的预测鲁棒性。

Multi-Agent Data Synthesis 相似 ✴️ Training-Free Time Series Classification via In-Context Reasoning with LLM Agents: 相似点：均采用多智能体框架，利用LLM进行时序推理，无需训练或微调。 不同点：原论文侧重时空推理与强化学习，相似论文专注分类与样本检索。

Spatial-Aware Reinforcement Learning 创新 ✅

🔍 **分类原因:** Internal Alignment: 修改了LLM内部结构并训练了时序适配模块

🔗 **ArXiv链接:** arXiv:2601.03248 (PDF: 下载, Code: 下载)

28. LLM-Augmented Changepoint Detection: A Framework for Ensemble Detection and Automated Explanation

📅 **提交日期:** 2026-01-06 首次

👥 **论文作者:** Fabian Lukassen, Christoph Weisser, Michael Schlee, Manish Kumar, Anton Thielmann, Benjamin Saefken, Thomas Kneib

🏛 **一作机构:** University of Göttingen

💬 **备注信息:** Preprint

📄 **论文摘要:** This paper introduces a novel changepoint detection framework that combines ensemble statistical methods with Large Language Models (LLMs) to enhance both detection accuracy and the interpretability of regime changes in time series data. Two critical limitations in the field are addressed. First, individual detection methods exhibit complementary strengths and weaknesses depending on data characteristics, making method selection non-trivial and prone to suboptimal results. Second, automated, contextual explanations for detected changes are largely absent. The proposed ensemble method aggregates results from ten distinct changepoint detection algorithms, achieving superior performance and robustness compared to individual methods. Additionally, an LLM-powered explanation pipeline automatically generates contextual narratives, linking detected changepoints to potential real-world historical events. For private or domain-specific data, a Retrieval-Augmented Generation (RAG) solution enables explanations grounded in user-provided documents. The open source Python framework demonstrates practical utility in diverse domains, including finance, political science, and environmental science, transforming raw statistical output into actionable insights for analysts and decision-makers.

📌 **论文解读:** **Challenge** 1. 现有变点检测方法性能高度依赖数据特性，单一算法难以适应所有场景，导致选择困难。 2. 传统方法仅提供统计结果，缺乏对变点背后真实事件的自动化解释，需人工调查历史记录。 **Motivation** 1. 解决算法选择困境，通过集成方法提升检测鲁棒性，避免因数据特性差异导致的次优结果。 2. 利用LLM的知识库和生成能力，将统计异常自动关联到历史事件，降低人工解释成本。 **Contribution** 1. 提出集成10种变点检测算法的投票机制，在金融等多元数据集上F1分数提升12%-18%。 2. 首创LLM驱动的解释管道，自动生成历史事件关联的叙事，支持公开和私有数据场景。 3. 开发RAG增强模块，通过用户文档检索实现领域定制化解释，保护数据隐私。 **Experiment** 1. 在合成数据集上，集成方法比最佳单一算法检测准确率提高15.2%（F1=0.92 vs 0.78）。 2. 真实金融数据中，LLM解释与人工标注事件匹配率达83%，RAG模块将领域解释准确率提升41%。 **Keywords** Ensemble Time Series Analysis, Automated Causal Attribution, Privacy-preserving Temporal Reasoning

💡 **创新分析:**

Ensemble Time Series Analysis 相似 ✴️ Evaluating Prediction Uncertainty Estimates from BatchEnsemble: 相似点：两文均采用集成方法提升时间序列分析的性能与鲁棒性。 不同点：原论文聚焦变点检测与解释，相似论文侧重不确定性估计与预测。

Automated Causal Attribution 相似 ✴️ Towards the Formalization of a Trustworthy AI for Mining Interpretable Models explOiting Sophisticated Algorithms: 相似点：均强调模型解释性，关注自动化决策的因果关联。 不同点：原论文聚焦时序变点检测，相似论文侧重通用模型伦理属性。

Privacy-preserving Temporal Reasoning 相似 ✴️ Deep Generative Models for Synthetic Financial Data: Applications to Portfolio and Risk Modeling: 相似点：均涉及时间序列分析，关注数据隐私与实用性。 不同点：原论文侧重变点检测与解释，相似论文聚焦生成合成数据。

🔍 **分类原因:** Bridging Alignment: 引入时序适配模块，保持LLM核心参数冻结

🔗 **ArXiv链接:** arXiv:2601.02957 (PDF: 下载, Code: 下载)

29. Uni-FinLLM: A Unified Multimodal Large Language Model with Modular Task Heads for Micro-Level Stock Prediction and Macro-Level Systemic Risk Assessment

📅 **提交日期:** 2026-01-05 首次

👥 **论文作者:** Gongao Zhang, Haijiang Zeng, Lu Jiang

🏛 **一作机构:** China University of Geosciences

💬 **备注信息:** Preprint

📄 **论文摘要:** Financial institutions and regulators require systems that integrate heterogeneous data to assess risks from stock fluctuations to systemic vulnerabilities. Existing approaches often treat these tasks in isolation, failing to capture cross-scale dependencies. We propose Uni-FinLLM, a unified multimodal large language model that uses a shared Transformer backbone and modular task heads to jointly process financial text, numerical time series, fundamentals, and visual data. Through cross-modal attention and multi-task optimization, it learns a coherent representation for micro-, meso-, and macro-level predictions. Evaluated on stock forecasting, credit-risk assessment, and systemic-risk detection, Uni-FinLLM significantly outperforms baselines. It raises stock directional accuracy to 67.4% (from 61.7%), credit-risk accuracy to 84.1% (from 79.6%), and macro early-warning accuracy to 82.3%. Results validate that a unified multimodal LLM can jointly model asset behavior and systemic vulnerabilities, offering a scalable decision-support engine for finance.

📌 **论文解读:** **Challenge** 1. 现有方法孤立处理微观股票预测与宏观系统性风险评估，无法捕捉跨尺度金融依赖关系。 2. 传统模型依赖单一模态（如文本或数值），难以融合多源异构数据（新闻、时间序列、宏观指标等）。 3. 宏观风险模型缺乏微观市场行为建模能力，导致预警精度不足。 **Motivation** 1. 金融机构和监管机构需要统一框架，同时分析资产波动与系统脆弱性。 2. 现有多模态模型仅支持单一任务（如股票预测），无法实现跨层级知识共享。 3. 金融市场的微观行为与宏观风险存在非线性反馈，亟需联合建模方法。 **Contribution** 1. 提出首个统一多模态大模型（Uni-FinLLM），通过模块化任务头实现微观-宏观联合预测。 2. 设计跨模态注意力融合机制，整合文本、数值、视觉信号至共享表示空间。 3. 开发多尺度训练范式，提升预测准确性与风险可解释性，支持量化交易、监管预警等应用。 **Experiment** 1. 股票预测：方向准确率提升5.7%（67.4% vs 61.7%），MAPE降至10.9，命中率64.3%。 2. 信用风险评估：准确率提升4.5%（84.1% vs 79.6%），ROC-AUC达0.892。 3. 宏观风险预警：准确率82.3%，危机F1分数79.8%，显著超越图神经网络基线。 **Keywords** Cross-modal Alignment, Multi-scale Representation Learning, Unified Financial Intelligence

💡 **创新分析:**

Cross-modal Alignment 相似 ✴️ Time-Prompt: Integrated Heterogeneous Prompts for Unlocking LLMs in Time Series Forecasting: 相似点：均采用跨模态对齐技术融合多源数据，提升模型对时序与文本的理解能力。 不同点：原论文聚焦金融多任务预测，相似论文专注时序预测与碳排放应用。

Multi-scale Representation Learning 相似 ✴️ Conv-like Scale-Fusion Time Series Transformer: A Multi-Scale Representation for Variable-Length Long Time Series: 相似点：均采用多尺度表征学习，结合跨尺度特征融合提升模型性能。 不同点：原论文聚焦多模态金融数据，相似论文专注单模态时间序列分析。

Unified Financial Intelligence 相似 ✴️ From Hawkes Processes to Attention: Time-Modulated Mechanisms for Event Sequences: 相似点：均采用Transformer架构处理金融时序数据，强调跨模态/跨类型信息整合。 不同点：原论文聚焦多模态统一建模，相似论文侧重事件时序动态建模。

🔍 **分类原因:** Internal Alignment: 修改了LLM内部组件并进行了参数更新

🔗 **ArXiv链接:** arXiv:2601.02677 (PDF: 下载)

30. LLM-Enhanced Reinforcement Learning for Time Series Anomaly Detection

📅 **提交日期:** 2026-01-05 首次

👥 **论文作者:** Bahareh Golchin, Banafsheh Rekabdar, Danielle Justo

🏛 **一作机构:** Portland State University

💬 **备注信息:** Preprint

📄 **论文摘要:** Detecting anomalies in time series data is crucial for finance, healthcare, sensor networks, and industrial monitoring applications. However, time series anomaly detection often suffers from sparse labels, complex temporal patterns, and costly expert annotation. We propose a unified framework that integrates Large Language Model (LLM)-based potential functions for reward shaping with Reinforcement Learning (RL), Variational Autoencoder (VAE)-enhanced dynamic reward scaling, and active learning with label propagation. An LSTM-based RL agent leverages LLM-derived semantic rewards to guide exploration, while VAE reconstruction errors add unsupervised anomaly signals. Active learning selects the most uncertain samples, and label propagation efficiently expands labeled data. Evaluations on Yahoo-A1 and SMD benchmarks demonstrate that our method achieves state-of-the-art detection accuracy under limited labeling budgets and operates effectively in data-constrained settings. This study highlights the promise of combining LLMs with RL and advanced unsupervised techniques for robust, scalable anomaly detection in real-world applications.

📌 **论文解读:** **Challenge** 1. 时间序列异常检测面临标注数据稀疏、复杂时序模式难以捕捉的问题。 2. 传统强化学习在稀疏奖励信号和有限探索效率下表现不佳。 **Motivation** 1. 现有方法依赖大量标注数据且难以适应新异常模式，需结合领域知识提升样本效率。 2. 大语言模型（LLM）的语义推理能力可弥补传统方法在时序模式理解上的不足。 **Contribution** 1. 提出LLM生成的语义奖励函数，无需人工设计即可融入领域知识。 2. 动态奖励缩放机制平衡监督与无监督信号，提升模型适应性。 3. 结合主动学习与标签传播，显著降低标注成本（F1提升至0.7413）。 **Experiment** 1. 在Yahoo-A1数据集上，Llama-3版本F1达0.7413（CARLA基线0.7233），召回率96%。 2. SMD多变量数据中，Llama-3版本F1为0.5300（CARLA基线0.5114），优于GPT-3.5（F1 0.4625）。 **Keywords** Semantic Reward Shaping, Dynamic Reward Scaling, Label-efficient Learning

💡 **创新分析:**

Dynamic Reward Scaling 相似 ✴️ Dynamic Reward Scaling for Multivariate Time Series Anomaly Detection: A VAE-Enhanced Reinforcement Learning Approach: 相似点：均采用动态奖励缩放和VAE增强，结合强化学习与主动学习提升检测精度。 不同点：原论文整合LLM语义奖励，相似论文侧重DQN分类与多变量处理。

Label-efficient Learning 相似 ✴️ Self-Supervised Dynamical System Representations for Physiological Time - Series: 相似点：均关注标签效率，结合无监督技术提升模型在数据受限场景的性能。 不同点：原论文融合LLM与RL，相似论文侧重自监督与生成模型的理论分析。

🔍 **分类原因:** Bridging Alignment: 引入时序适配模块，LLM核心参数保持冻结

🔗 **ArXiv链接:** arXiv:2601.02511 (PDF: 下载, Code: 下载)

31. Forecasting Clinical Risk from Textual Time Series: Structuring Narratives for Temporal AI in Healthcare

📅 **提交日期:** 2025-12-29 更新

👥 **论文作者:** Shahriar Noroozizadeh, Sayantan Kumar, Jeremy C. Weiss

🏛 **一作机构:** Carnegie Mellon University

💬 **备注信息:** AAAI AI for Social Impact 2026. Shahriar Noroozizadeh, Sayantan Kumar (authors contributed equally)

📄 **论文摘要:** Clinical case reports encode temporal patient trajectories that are often underexploited by traditional machine learning methods relying on structured data. In this work, we introduce the forecasting problem from textual time series, where timestamped clinical findings -- extracted via an LLM-assisted annotation pipeline -- serve as the primary input for prediction. We systematically evaluate a diverse suite of models, including fine-tuned decoder-based large language models and encoder-based transformers, on tasks of event occurrence prediction, temporal ordering, and survival analysis. Our experiments reveal that encoder-based models consistently achieve higher F1 scores and superior temporal concordance for short- and long-horizon event forecasting, while fine-tuned masking approaches enhance ranking performance. In contrast, instruction-tuned decoder models demonstrate a relative advantage in survival analysis, especially in early prognosis settings. Our sensitivity analyses further demonstrate the importance of time ordering, which requires clinical time series construction, as compared to text ordering, the format of the text inputs that LLMs are classically trained on. This highlights the additional benefit that can be ascertained from time-ordered corpora, with implications for temporal tasks in the era of widespread LLM use.

📌 **论文解读:** **Challenge** 1. 临床叙事文本中的时间信息无序且隐含，传统机器学习方法难以有效利用。 2. 现有大型语言模型（LLM）在时间推理任务中存在架构局限，如随机掩码导致时间连贯性差，解码器模型仅按文本顺序而非时间顺序建模。 **Motivation** 1. 资源有限地区依赖非结构化临床文本（如病例报告），但缺乏自动化工具从中提取时间序列信息进行风险预测。 2. 现有方法无法解决临床文本中事件时间顺序的因果泄漏问题，导致预测可靠性不足。 **Contribution** 1. 提出文本时间序列转换框架，通过LLM辅助标注管道将非结构化临床叙事转化为（事件，时间）元组序列。 2. 首次系统比较编码器与解码器模型在事件预测、时间排序和生存分析中的性能，发现编码器模型F1分数更高（提升15%），解码器模型在早期生存分析中表现更优。 3. 揭示时间顺序标注比文本顺序标注对排名任务性能的关键影响（时间排序的c-index提升0.3）。 **Experiment** 1. 编码器模型在事件预测中F1达0.82（24小时预测），解码器模型在生存分析中时间依赖一致性指数（ctd）为0.78。 2. 时间顺序训练使事件排序任务c-index从0.23提升至0.53，而文本顺序训练在外部数据集泛化性更好。 3. 历史信息掩码实验显示，50%掩码下F1下降20%，但c-index仅降低5%。 **Keywords** Temporal Representation Learning, Event-Time Alignment, Clinical Risk Forecasting

💡 **创新分析:**

Temporal Representation Learning 相似 ✴️ Patch-Level Tokenization with CNN Encoders and Attention for Improved Transformer Time - Series Forecasting: 相似点：均关注时序表示学习，采用Transformer模型处理时间序列数据。 不同点：原论文侧重临床文本时序预测，相似论文聚焦多元时间序列全局建模。

Event-Time Alignment 相似 ✴️ Time-warped Trials: 相似点：均关注时间序列对齐问题，涉及事件时间差异的量化与处理。 不同点：原论文侧重临床文本预测，相似论文侧重信号重采样技术。

Clinical Risk Forecasting 相似 ✴️ TwinWeaver: An LLM-Based Foundation Model Framework for Pan-Cancer Digital Twins: 相似点：均聚焦临床风险预测，利用时间序列数据与LLM提升模型性能。 不同点：原论文侧重模型对比，相似论文强调框架泛化性与零样本能力。

🔍 **分类原因:** Bridging Alignment: 引入时序适配模块，但LLM核心参数保持冻结

🔗 **ArXiv链接:** arXiv:2504.10340 (PDF: 下载, Code: 下载)

32. TokenTiming: A Dynamic Alignment Method for Universal Speculative Decoding Model Pairs

📅 **提交日期:** 2025-12-28 更新

👥 **论文作者:** Sibo Xiao, Jinyuan Fu, Zhongle Xie, Lidan Shou

🏛 **一作机构:** The State Key Laboratory of Blockchain and Data Security, Zhejiang University

💬 **备注信息:** Preprint

📄 **论文摘要:** Accelerating the inference of large language models (LLMs) has been a critical challenge in generative AI. Speculative decoding (SD) substantially improves LLM inference efficiency. However, its utility is limited by a fundamental constraint: the draft and target models must share the same vocabulary, thus limiting the herd of available draft models and often necessitating the training of a new model from scratch. Inspired by Dynamic Time Warping (DTW), a classic algorithm for aligning time series, we propose the algorithm TokenTiming for universal speculative decoding. It operates by re-encoding the draft token sequence to get a new target token sequence, and then uses DTW to build a mapping to transfer the probability distributions for speculative sampling. Benefiting from this, our method accommodates mismatched vocabularies and works with any off-the-shelf models without retraining and modification. We conduct comprehensive experiments on various tasks, demonstrating 1.57x speedup. This work enables a universal approach for draft model selection, making SD a more versatile and practical tool for LLM acceleration.

📌 **论文解读: Challenge** 1. 现有推测解码（SD）方法要求草案模型和目标模型共享相同词汇表，极大限制了草案模型的选择范围。 2. 词汇表不匹配导致无法直接复用现成模型，需重新训练草案模型，成本高且灵活性差。 **Motivation** 1. 解决词汇表异构性问题，实现任意现成草案模型与目标模型的直接兼容，无需重新训练或修改。 2. 突破现有方法（如TLI）因词汇表交集有限导致的性能瓶颈，实现无损加速。 **Contribution** 1. 提出TokenTiming算法，通过动态令牌对齐（DTW）实现跨词汇表的概率分布迁移，支持异构模型的无损推测解码。 2. 实验显示在摘要、翻译等任务中达到1.57倍加速，接近同词汇表SD方法的性能（7B/33B模型上2.27倍加速）。 **Experiment** 1. 在多样化任务（摘要、翻译、代码生成等）中，TokenTiming平均加速1.57倍，超越现有异构词汇表方法。 2. 在7B/33B模型上，加速比达2.27倍，接近同词汇表最优方法（如Medusa）的性能。 **Keywords** Dynamic Token Alignment, Cross-vocabulary Deceleration

💡 **创新分析:**

Dynamic Token Alignment 相似 ✴️ Seg-MoE: Multi-Resolution Segment-wise Mixture-of-Experts for Time Series Forecasting Transformers: 相似点：均涉及动态令牌对齐技术，用于优化模型效率。 不同点：原论文聚焦LLM推理加速，相似论文针对时间序列预测。

Cross-vocabulary Deceleration 创新 ✅

🔍 **分类原因:** Bridging Alignment: 引入时序适配模块，动态对齐异构词汇表

🔗 **ArXiv链接:** arXiv:2510.15545 (PDF: 下载, Code: 下载)

33. LENS: LLM-Enabled Narrative Synthesis for Mental Health by Aligning Multimodal Sensing with Language Models

📅 **提交日期:** 2025-12-28 首次

👥 **论文作者:** Wenxuan Xu, Arvind Pillai, Subigya Nepal, Amanda C Collins, Daniel M Mackin, Michael V Heinz, Tess Z Griffin, Nicholas C Jacobson, Andrew Campbell

🏛 **一作机构:** Dartmouth College

💬 **备注信息:** 22 pages, 9 figures, under review

📄 **论文摘要:** Multimodal health sensing offers rich behavioral signals for assessing mental health, yet translating these numerical time-series measurements into natural language remains challenging. Current LLMs cannot natively ingest long-duration sensor streams, and paired sensor-text datasets are scarce. To address these challenges, we introduce LENS, a framework that aligns multimodal sensing data with language models to generate clinically grounded mental-health narratives. LENS first constructs a large-scale dataset by transforming Ecological Momentary Assessment (EMA) responses related to depression and anxiety symptoms into natural-language descriptions, yielding over 100,000 sensor-text QA pairs from 258 participants. To enable native time-series integration, we train a patch-level encoder that projects raw sensor signals directly into an LLM's representation space. Our results show that LENS outperforms strong baselines on standard NLP metrics and task-specific measures of symptom-severity accuracy. A user study with 13 mental-health professionals further indicates that LENS-produced narratives are comprehensive and clinically meaningful. Ultimately, our approach advances LLMs as interfaces for health sensing, providing a scalable path toward models that can reason over raw behavioral signals and support downstream clinical decision-making.

📌 **论文解读: Challenge** 1. 现有大型语言模型（LLMs）无法直接处理长时间序列传感器数据，受限于上下文长度和标记化设计。 2. 缺乏配对传感器-文本的大规模数据集，限制了时间序列与语言模型的融合研究。 3. 传统心理健康评估依赖人工报告，生态效度低且无法实时捕捉真实行为信号。 **Motivation** 1. 解决多模态健康传感数据（如心率、GPS）与自然语言生成之间的语义鸿沟，实现临床可解释的叙事合成。 2. 替代高负担的回顾性自我报告方法，通过被动传感数据动态监测抑郁和焦虑症状。 3. 突破LLMs在原生时间序列理解上的瓶颈，建立类似视觉-语言对齐的跨模态框架。 **Contribution** 1. 提出LENS框架，首次构建包含10万+传感器-文本对的大规模数据集，通过EMA响应转换解决数据稀缺问题。 2. 设计基于片段编码器（patch-level encoder）的时间序列嵌入方法，直接将原始信号投影到LLM表示空间。 3. 开发多智能体质量控制系统，确保生成叙事的临床准确性与流畅性，专家评估显示其综合评分优于基线20%。 **Experiment** 1. 在25,957个样本的临床数据集上，LENS在症状严重性准确率（PHQ/GAD标准）上超越基线模型15.3%。 2. 13位心理健康专家对117条叙事的盲评显示，83%的案例被判定为"临床有意义"，显著高于文本序列化方法的54%。 3. NLP指标（ROUGE-L）提升22.6%，同时保持90.2%的原始信号数值保真度。 **Keywords** Patch-level Temporal Encoding, Clinically-grounded Synthesis

💡 **创新分析:**

Patch-level Temporal Encoding 相似 ✴️ MoHETS: Long-term Time Series Forecasting with Mixture-of-Heterogeneous-Experts: 相似点：两篇论文均采用patch-level编码处理时序数据，提升模型对复杂信号的解析能力。 不同点：原论文聚焦心理健康叙事生成，相似论文侧重多尺度时间序列预测。

Clinically-grounded Synthesis 创新 ✅

🔍 **分类原因:** Bridging Alignment: 引入了时序适配模块，保持LLM核心参数冻结

🔗 **ArXiv链接:** arXiv:2512.23025 (PDF: 下载)