Financial TimeSeries Reasoning Benchmarks at Scale

Malgorzata Gwiazda[1], **Yifu Cai**[2], **Mononito Goswami**[2], **and Artur Dubrawski**[2]

[1]Technical University of Munich, Munich, Germany
malgorzata.gwiazda@tum.de

[2]Auton Lab, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213
{yifuc, mgoswami, awd}@cs.cmu.edu

Abstract

We introduce `TimeSeriesExamAgent`, a scalable and domain-agnostic framework for automatically generating and validating time series reasoning benchmarks. Existing benchmarks lack scalability, are limited to a few specific domains, while building them remains labor intensive. Automated solutions for benchmark creation have been proposed, but these typically rely on a single-step generation process without verification, leading to lower-quality exams. Our framework addresses these limitations by enabling stakeholders—such as financial institutions with highly confidential data—to easily construct high-quality, domain-specific benchmarks from their own private datasets. A domain expert provides a dataset, a natural language description, and a simple data-loading method. The agent then orchestrates the generation pipeline, including creating question templates, robustness verification from multiple perspectives, and iterative refinement. We demonstrate the framework on financial datasets and evaluate multiple state-of-the-art language models on the generated benchmarks. Empirically, we demonstrate that the framework produces domain-agnostic benchmarks whose diversity matches human-generated counterparts, and our evaluation of several Large Language Models shows that accuracy remains limited, underscoring open challenges in time-series reasoning.

1 Introduction

Many recent works have applied Large Language Models (LLMs) to time series analysis tasks such as forecasting, anomaly detection, and classification [1, 2, 3, 4, 5, 6]. More recently, attention has shifted to evaluating the reasoning capabilities of LLMs in time series tasks. These evaluations are typically framed in two ways: 1) contextualized traditional tasks such as forecasting, but with added contextual information (e.g., providing a macro economic scenario before a prediction) [7, 8, 9, 10, 11], and 2) reasoning and understanding tasks that directly probe concepts in time series (e.g., "what kind of trend does the following series exhibit?") [12, 13].

However, existing benchmarks have clear limitations. Contextualized tasks remain close to traditional metrics (e.g., mean-squared-error for forecasting) without testing deeper reasoning, while reasoning-style benchmarks often focus only on simple properties like trend or seasonality. In practice, real-world domains such as healthcare require more complex reasoning, where tasks like diagnosis naturally combine anomaly detection, classification, and domain knowledge. Curation is another challenge. Annotation or template-based benchmarks are labor-intensive, while LLM-based augmentation often lacks diversity because it simply expands existing datasets. As a result, building specialized, domain-specific benchmarks remains difficult and time-consuming.

This challenge is especially pronounced in **finance**. Financial datasets are both highly proprietary and highly specialized: for example, tasks in different asset classes often require nuanced reasoning

that combines statistical patterns with market microstructure and domain knowledge. Unlike open domains where public benchmarks are available, stakeholders in finance cannot easily share their datasets due to privacy, confidentiality, and competitive concerns. At the same time, they need ways to rigorously evaluate whether generative models can reason about specialized topics without exposing sensitive data. This creates an urgent need for customizable, automated benchmark generation tailored to private financial datasets.

Inspired by recent agent-based approaches in other domains [14, 15], we propose `TimeSeriesExamAgent`, a pipeline that (1) generates domain-specific multiple-choice questions on time-series data, (2) scales efficiently, and (3) ensures reliable ground truth through iterative verification. We also evaluate four state-of-the-art LLMs on a benchmark of 290 automatically generated questions. Our results show that many models struggle on highly complex tasks that require combining quantitative analysis with financial domain knowledge. For brevity, we provide detailed related work in Appendix A.

2 TimeSeriesExamAgent

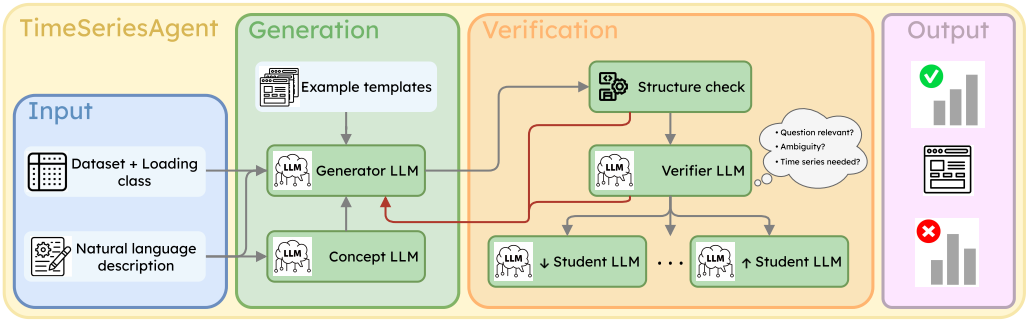


Figure 1: TimeSeriesExamAgent architecture. The user provides exam-making instructions and a custom dataset with minimal loading code. Agent outputs question templates – Python functions generated by a generator LLM and filtered through three progressive stages of verification (syntax and output format check, validation by LLM judge, capability-aligned filtering). Arrows denote data flow, red ones show direction for rejected templates.

In this section, we introduce `TimeSeriesExamAgent`, a multi-agent framework that combines planning, generation, and verification to enable automatic benchmark construction. An overview is shown in Fig. 1. The Generation Agent takes as input a description of the natural language task T and a data set D. The description T may include user guidelines for generation, contextual information about the dataset, or other relevant instructions. For convenience, we denote each sample in D as (x_i, z_i), where $x_i \in \mathbb{R}^{n \times d}$ is a time series with n observations and d variables, and z_i is an auxiliary array containing metadata or labels related to the series. The user provides a dataset class D that supports basic operations such as querying the i-th sample.

Generation We generate question templates instead of samples directly, as shown in Fig. 2. Templates offer two advantages: they are scalable, and their abstraction adds an extra layer of robustness. By relying on structured, rule-based generation rather than manual inputs, they reduce the chance of human errors or inconsistencies. Our generator LLM produces a predefined number of templates, each implemented as a Python function. A template contains a formatted string for the question and options, together with parameters that control how many questions to generate. For each question, the template samples a pair (x_i, z_i) from the dataset D and applies a rule-based calculation to determine the correct answer from the time series. For example, in a trend-detection template, the function computes the linear trend coefficient of x_i and selects "Yes, there is a linear trend" if the coefficient exceeds a specified threshold. In addition to such signal-derived logic, templates can also utilize the auxiliary property z_i, effectively transforming classification problems into question–answer form. For instance, if equity data is labeled according to positive or negative earnings surprises, the template can convert these labels into multiple-choice options. Each generated sample consists of the question, its options, the correct answer, and one or more associated time series represented as numerical

values. We provide a breakdown of the Generation Agent and its prompt in Appendix B. An example template is also provided.

Verification We observe that LLM-based generation frequently produces errors or irrelevant outputs, motivating the need for a structured verification process. We propose a multistage verification process to check the accuracy and relevance of each template. If a template fails at any stage, it is returned to the generation agent with feedback. The generation is iterative with a maximum of three attempts, after which the ongoing template is discarded to avoid excessive context length and cost from repeated failures.

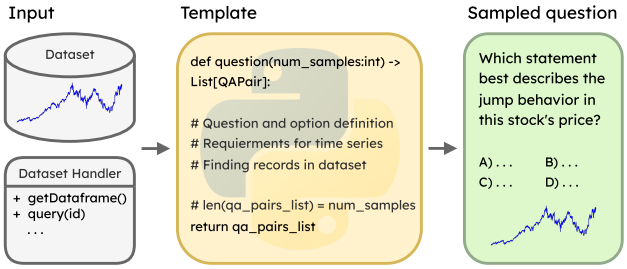


Figure 2: Question generation process: With information about dataset, TimeSeriesAgent generates question template in a form of Python functions. The created function can be called to get arbitrary number of question samples.

Structure verification We check whether the generated template can be executed successfully. We execute the generated template $k = 5$ times; if there are failures, the error message is returned as a feedback.

Content verification Certain aspects of quality control are particularly well-suited for LLM-as-a-judge evaluation. For example, verifying that a question is grammatically correct, free of ambiguity or bias, and genuinely answerable from the accompanying time series can be effectively handled by an LLM. To this end, we use an LLM verifier to assess the validty of each template. A quantitative score is given, and we set a threshold for rejection. If the verifier raises any rejection, its explanation is treated similarly to a structural error and the template is regenerated. We provide the detailed prompt in Appendix C.

Capability-Aligned Filtering To detect templates that generate overly simple or irrelevant exams, we evaluate them using a set of test-taking LLMs with varying capabilities. This approach is inspired by educational theory, particularly the expertise reversal effect [16]. A template is discarded if weaker LLMs achieve higher average accuracy than stronger models, as this typically indicates that the template is flawed or noisy rather than genuinely discriminative. Templates are retained if performance scales with model capability– or if all models perform poorly, since such questions may still capture genuine difficulty. We provide hyper-parameters in Appendix F and other design specifics in Appendix D

3 Experimental Setup, Results and Discussion

First, we generate one example using real-world stock price data retrieved from Yahoo Finance. First, we generate one exam for the three real world stock price datasets from yahoo finance [17]. In total, we have 290 samples for YFinance, where we sample 5 instances per template.

Model	YFinance
gpt-4o [18]	**0.586**
o3-mini [19]	0.555
Qwen2.5-VL-Instruct [20]	0.572
Gemma-3-27b-it [21]	0.534

Table 1: Comparative performance of four vision–language models across YFinance dataset, measured with accuracy. The results reveal gpt-4o as the best performer. Nonetheless, all models achieve less than 60 mean accuracy, underscoring the difficulty of time-series reasoning for current VLMs. The evaluation protocol is provided in Appendix G

We select candidate models to cover a diverse range of performance levels, as indicated by the OpenVLM Leaderboard [22]. In Table 1, we find that strong, general-purpose model such as `gpt-4o` perform well on finance-related questions.

To further evaluate our benchmark, we compare multiple metrics on questions generated from the dataset with those in `MTBench` [7], a timeseries benchmark which also contains multiple choice questions from the domain of finance. We also used FinMME, a financial benchmark covering 18 different finance domains, which was partially human-annotated and expert-validated. The goal is to demonstrate that our framework achieves comparable diversity without requiring manual template curation. We picked random 50 question samples from each benchmark and calculated the distances for every possible pair within the set. We used the `Qwen/Qwen3-0.6B` sentence transformer model to extract embeddings, as it achieved the forth-best performance among all models on the Hugging Face MTEB leaderboard.

Benchmark Dataset	**Mean ± Std**	
	Embedding	Normalized Levenshtein
MTBench	0.454 ± 0.056	0.490 ± 0.015
FinMME	0.582 ± 0.082	0.432 ± 0.079
TimeSeriesExamAgent (ours)	$\mathbf{0.590 \pm 0.095}$	$\mathbf{0.540 \pm 0.052}$

Table 2: Question diversity comparison using embedding and normalized Levenshtein distance.

As shown in Table 2, benchmark generated by our framework shows a diversity comparable to MTBench and FinMME. This indicates that the proposed framework is able to capture a wide range of expressions without manual intervention, supporting its scalability and adaptability to other domains.

We also employed G-Eval, a probabilistic LLM-as-a-judge framework [23]. An LLM is used to evaluate the relevance of each question, assigning a score between 0 and 1 to indicate how well it meets the specified criteria. Results are presented in Table 3. We provide the detailed G-Eval prompt in Appendix E.

Dataset	**Mean Result**			
	Specificity	Unambiguity	Domain Relevance	Answerability
MTBench	0.789	0.883	**0.999**	0.921
FinMME	0.820	0.865	0.816	0.883
TimeSeriesExamAgent (ours)	**0.976**	**0.948**	0.977	**0.926**

Table 3: Question diversity comparison using G-Eval framework.

4 Limitations and Conclusions

In this work, we present a scalable, domain-specific framework for the automatic generation of time-series benchmarks, enabling the creation of high-quality, large-scale evaluation datasets while minimizing the need for labor-intensive human annotation.

A limitation of this study is that the quality of the generated exams depends on the quality and coverage of the time series dataset. Additionally, domain specialists must provide carefully crafted prompts. The study also lacks statistical significance testing when comparing results.

Although we evaluate only on time-series and multiple-choice questions, the pipeline generalizes beyond both modalities; exploring these settings is left to future work. We also plan to explore privacy-aware question generation and to validate exam quality by training time-series–text alignment models and testing their transfer performance on other established reasoning benchmarks [8]. Finally, there is growing attention on building time-series agentic frameworks [24, 25]. Enabling these frameworks to write code in order to answer our benchmark questions would provide valuable insights to the community.

Acknowledgments

This work has been partially supported by the National Science Foundation (awards 2427948, 2406231 and 2530752) and National Institutes of Health (awards R01NR013912 and R01NS124642) and with the gift from Dr. Chirag Nagpal.

We plan to make the code publicly available in the recent future.

References

[1] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

[2] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

[3] Defu Cao, Furong Jia, Sercan O Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. Tempo: Prompt-based generative pre-trained transformer for time series forecasting. *arXiv preprint arXiv:2310.04948*, 2023.

[4] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*, 2023.

[5] Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36:43322–43355, 2023.

[6] Nina Żukowska, Mononito Goswami, Michał Wiliński, Willa Potosnak, and Artur Dubrawski. Towards long-context time series foundation models. *arXiv preprint arXiv:2409.13530*, 2024.

[7] Jialin Chen, Aosong Feng, Ziyu Zhao, Juan Garza, Gaukhar Nurbek, Cheng Qin, Ali Maatouk, Leandros Tassiulas, Yifeng Gao, and Rex Ying. Mtbench: A multimodal time series benchmark for temporal reasoning and question answering. *arXiv preprint arXiv:2503.16858*, 2025.

[8] Yaxuan Kong, Yiyuan Yang, Yoontae Hwang, Wenjie Du, Stefan Zohren, Zhangyang Wang, Ming Jin, and Qingsong Wen. Time-mqa: Time series multi-task question answering with context enhancement. *arXiv preprint arXiv:2503.01875*, 2025.

[9] Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Prabhakar Kamarthi, Aditya Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, et al. Time-mmd: Multi-domain multimodal dataset for time series analysis. *Advances in Neural Information Processing Systems*, 37:77888–77933, 2024.

[10] Jungwoo Oh, Gyubok Lee, Seongsu Bae, Joon-myoung Kwon, and Edward Choi. Ecg-qa: A comprehensive question answering dataset combined with electrocardiogram. *Advances in Neural Information Processing Systems*, 36:66277–66288, 2023.

[11] Xu Wang, Jiaju Kang, Puyu Han, Yubao Zhao, Qian Liu, Liwenfei He, Lingqiong Zhang, Lingyun Dai, Yongcheng Wang, and Jie Tao. Ecg-expert-qa: A benchmark for evaluating medical large language models in heart disease diagnosis. *arXiv preprint arXiv:2502.17475*, 2025.

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

[13] Willa Potosnak, Cristian Challu, Mononito Goswami, Kin G. Olivares, Michał Wiliński, Nina Żukowska, and Artur Dubrawski. Investigating compositional reasoning in time series foundation models, 2025.

[14] Natasha Butt, Varun Chandrasekaran, Neel Joshi, Besmira Nushi, and Vidhisha Balachandran. Benchagents: Automated benchmark creation with agent interaction. *arXiv preprint arXiv:2410.22584*, 2024.

[15] Gauthier Guinet, Behrooz Omidvar-Tehrani, Anoop Deoras, and Laurent Callot. Automated evaluation of retrieval-augmented language models with task-specific exam generation. *arXiv preprint arXiv:2405.13622*, 2024.

[16] Slava Kalyuga. Expertise reversal effect and its implications for learner-tailored instruction. *Educational psychology review*, 19(4):509–539, 2007.

[17] Ranan Roussi. yfinance: Yahoo! finance market data downloader. `https://github.com/ranaroussi/yfinance`, 2017. Accessed: 2025-08-22.

[18] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[19] OpenAI. Openai o3-mini system card. `https://openai.com/index/o3-mini-system-card/`, January 2025. Accessed: 2025-08-22.

[20] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[21] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[22] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201, 2024.

[23] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.

[24] Yifu Cai, Xinyu Li, Mononito Goswami, Michał Wiliński, Gus Welter, and Artur Dubrawski. Timeseriesgym: A scalable benchmark for (time series) machine learning engineering agents. *arXiv preprint arXiv:2505.13291*, 2025.

[25] Wen Ye, Wei Yang, Defu Cao, Yizhou Zhang, Lumingyuan Tang, Jie Cai, and Yan Liu. Domain-oriented time series inference agents for reasoning and automated analysis. *arXiv preprint arXiv:2410.04047*, 2024.

[26] Yilin Wang, Peixuan Lei, Jie Song, Yuzhe Hao, Tao Chen, Yuxuan Zhang, Lei Jia, Yuanxiang Li, and Zhongyu Wei. Itformer: Bridging time series and natural language for multi-modal qa with large-scale multitask dataset. *arXiv preprint arXiv:2506.20093*, 2025.

[27] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*, 2021.

[28] Dong Shu, Haoyang Yuan, Yuchen Wang, Yanguang Liu, Huopu Zhang, Haiyan Zhao, and Mengnan Du. Finchart-bench: Benchmarking financial chart comprehension in vision-language models. *arXiv preprint arXiv:2507.14823*, 2025.

[29] Junyu Luo, Zhizhuo Kou, Liming Yang, Xiao Luo, Jinsheng Huang, Zhiping Xiao, Jingshu Peng, Chengzhong Liu, Jiaming Ji, Xuanzhe Liu, Sirui Han, Ming Zhang, and Yike Guo. Finmme: Benchmark dataset for financial multi-modal reasoning evaluation, 2025.

[30] Wanying Wang, Zeyu Ma, Pengfei Liu, and Mingang Chen. Testagent: A framework for domain-adaptive evaluation of llms via dynamic benchmark construction and exploratory interaction. *arXiv preprint arXiv:2410.11507*, 2024.

[31] Mike A Merrill, Mingtian Tan, Vinayak Gupta, Tom Hartvigsen, and Tim Althoff. Language models still struggle to zero-shot reason about time series. *arXiv preprint arXiv:2404.11757*, 2024.

[32] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[33] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[34] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[35] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020.

[36] Jesse Mu, Xiang Lisa Li, and Noah Goodman. Learning to compress prompts with gist tokens, 2024.

[37] Qingyue Wang, Yanhe Fu, Yanan Cao, Shuai Wang, Zhiliang Tian, and Liang Ding. Recursively summarizing enables long-term dialogue memory in large language models, 2025.

[38] Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, et al. Openhands: An open platform for ai software developers as generalist agents. *arXiv preprint arXiv:2407.16741*, 2024.

A Related Work

Time series benchmarks The task of creating time series reasoning benchmarks is challenging. Existing benchmarks are either domain-agnostic, or limited to a specific domains with high quality datasets. For example, `TimeSeriesExam` [12] introduced over 700 multiple-choice questions to evaluate five general reasoning skills, but its questions primarily assess signal properties (e.g. trend, cyclicity, stationarity) and lack the contextual depth needed for real-world applications. Domain-specific benchmarks address this gap but have limited scope and poor extensibility, since their curation often relies on templates. For instance, ECG-QA [10] and ECG-Expert-QA [11] focus on ECG interpretation, while EngineMT-QA [26] targets industrial settings. Automatic benchmark generation offers a scalable alternative but raises concerns about quality and diversity of generated questions. Without extensive verification, LLM-generated questions often require heavy manual curation [8, 9], which is both difficult and time-consuming—undermining the main advantage of automation.

Financial reasoning benchmarks Reasoning tasks play a crucial role in financial benchmarks. Such tasks may be oriented around tabular data understanding, extraction of information from charts, or analysis of stock prices. In the FinQA [27] dataset, the questions are focused on financial numerical data understanding, but does not extend to chart interpretation This gap is filled by FinChart-Bench [28] introduces tasks centered on extracting and reasoning over chart information. While many charts involve aspects of temporal change, only a subset suits time series understanding task. the FinMME [29] benchmark covers multiple modalities -— charts, numerical data, and text —-through MCQA questions, including time series tasks, though its construction required over 800 hours of manual annotation.

Finance is frequently incorporated as one of the domains in multi-domain time series benchmarks. MTBench [7] includes question–answer pairs in both the weather and finance domains. It links trading-related articles with relevant stock prices, creating a real-world, grounded multiple-choice

Title	Multi-Domain	Curation	# Samples	Skill type		
		Fully Automatic		P	R	PS
Time-MQA [8]	✓	✗	200,000	✓	✓	✓
TimeSeriesExam [12]	✗	✗	763	✓	✓	✗
Time-MMD [9]	✓	✗	17,113	✓	✗	✗
MT-Bench [7]	✓	✓	22,000	✓	✓	✗
ECG-QA [10]	✗	✗	414,348	✓	✓	✗
TimeSeriesExamAgent (ours)	✓	✓	600+	✓	✓	✓

Table 4: Overview of time series and multimodal datasets with curation and skill types (P – Prediction, R – Reasoning, PS – Practical skills (tasks beyond classification and reasoning such as performing calculations and applying formulas). TimeSeriesExamAgent is universal – tailored to user's needs and with advanced automatic verifications.

question dataset. However, MTBench remains difficult to scale and suffers from limited question diversity.

Agents for benchmark creation An AI agent is an autonomous system that can observe its environment, reason about possible actions, and act toward achieving a goal. In LLM-based settings, the language model often provides the reasoning or planning layer that guides the agent's decisions. Recent work has shown success in using agents for automatic benchmark creation. Most solutions adopt a multi-agent pipeline with planning, generation, validation, and evaluation modules [14]. For instance, [30] integrates exploratory evaluation using reinforcement learning, while [14] takes a natural language task description as input. However, most of these approaches are not tailored to time series and struggle to generate questions conditioned on numeric data. One recent solution does incorporate time series but is limited to single-step design and lacks extensive verification [31].

B Generation Agent Workflow

We rely on two stages of generation for the templates: planning and generating, inspired by the chain-of-thought (CoT) prompting[32].

Generation planning To provide a relevant and diverse set of templates, we rely on a comprehensive list of domain-specific concepts. There are several ways our pipeline generates a list of concepts:

1. LLM generation: User guidelines and dataset descriptions are provided as input to an LLM, which proposes the concepts.
2. Web Search: We provide the option for generator LLM obtain concepts through web search.
3. Retrieval Augmented Generation: As an option, the user could also provide a relevant file from which the LLM reads and generates concepts[33].

Template generation As input to our generator, the following components are provided:

- User-provided guidelines: a document containing the user's goal or specific requirements,
- Dataset description: a list of columns and example values with ranges from the dataset, with a short usage example,
- List of concepts: generated in previous step. For each template, our pipeline will choose a concept at random to ensure diversity.
- Example templates[Optional]: user-provided few-shot examples presenting required structural elements [34].

B.1 Generation Prompt

```
Here is the goal of the exam questions:
{user_info_text}

Here are sample concepts on which you can base your question generation:
{concept_conversation}

Use the concept numbered {concept_no} from the list to guide the design of
    your question template.

Here is the description of the dataset you will use to generate the
    question:
{dataset_describe}

In your template, use the provided ‘user_dataset‘ object. Use its ‘query(
    index)‘ method to load relevant time series data.

Do not select time series randomly. First, formulate the question, and then
     choose a time series that fits its logic and reasoning needs.

Generate one function-based question template now.
```

B.2 Example of Question Template

```
def question_6(num_samples, verbose=False):
    hyperparameters = {
        "min_trend_days": 20,
        "max_series_length": 3000,
        "trend_strength_threshold": 0.7,
        "momentum_window": 10,
    }

    question = "Analyzing the price movements of {ticker} over the given
    time period, does the price trend demonstrate strong momentum and
    sustainability, or does it show signs of weakness and potential
    reversal?"

    options = [
        "The trend shows strong momentum with consistent directional
    movement and minimal pullbacks, suggesting the trend is likely to
    continue.",
        "The trend shows signs of weakness with frequent reversals and
    inconsistent momentum, suggesting a potential trend change.",
        "The trend shows mixed signals with alternating periods of strength
     and weakness, making direction unclear.",
        "The price movement shows no clear trend pattern, indicating a
    ranging or sideways market."
    ]

    def calculate_trend_strength(prices):
        if len(prices) < hyperparameters["min_trend_days"]:
            return None, None

        returns = np.diff(prices) / prices[:-1]

        # Calculate momentum consistency
        positive_days = np.sum(returns > 0)
        negative_days = np.sum(returns < 0)
        total_days = len(returns)
```

```
    directional_consistency = max(positive_days, negative_days) /
total_days

    # Calculate average magnitude of moves
    avg_abs_return = np.mean(np.abs(returns))

    # Calculate trend persistence (consecutive moves in same direction)
    consecutive_moves = []
    current_streak = 1
    for i in range(1, len(returns)):
        if np.sign(returns[i]) == np.sign(returns[i-1]):
            current_streak += 1
        else:
            consecutive_moves.append(current_streak)
            current_streak = 1
    consecutive_moves.append(current_streak)

    avg_streak = np.mean(consecutive_moves)
    max_streak = max(consecutive_moves)

    # Determine overall trend direction
    overall_return = (prices[-1] - prices[0]) / prices[0]
    trend_direction = "up" if overall_return > 0 else "down"

    return {
        "directional_consistency": directional_consistency,
        "avg_abs_return": avg_abs_return,
        "avg_streak": avg_streak,
        "max_streak": max_streak,
        "overall_return": abs(overall_return),
        "trend_direction": trend_direction
    }, returns

qa_pairs = []
df = user_dataset.get_dataframe()

attempted_tickers = set()

while len(qa_pairs) < num_samples:
    if verbose:
        print(f"[Question 6] Generating question {len(qa_pairs)} / {
num_samples}")

    # Select a ticker that hasn't been attempted
    available_tickers = [i for i in df.index if i not in
attempted_tickers]
    if not available_tickers:
        break

    ticker_id = random.choice(available_tickers)
    attempted_tickers.add(ticker_id)

    ticker = df.loc[ticker_id, 'ticker']
    prices = user_dataset.query(ticker_id)

    if len(prices) < hyperparameters["min_trend_days"]:
        continue

    # Limit series length
```

```
    if len(prices) > hyperparameters["max_series_length"]:
        start_idx = random.randint(0, len(prices) - hyperparameters["
max_series_length"])
        prices = prices[start_idx:start_idx + hyperparameters["
max_series_length"]]

    # Select a subset for analysis (to make question more focused)
    analysis_length = min(len(prices), random.randint(50, 200))
    start_idx = random.randint(0, len(prices) - analysis_length)
    analysis_prices = prices[start_idx:start_idx + analysis_length]

    trend_metrics, returns = calculate_trend_strength(analysis_prices)
    if trend_metrics is None:
        continue

    # Determine answer based on trend strength metrics
    strength_score = (
        trend_metrics["directional_consistency"] * 0.4 +
        min(trend_metrics["avg_streak"] / 5, 1.0) * 0.3 +
        min(trend_metrics["overall_return"] * 10, 1.0) * 0.3
    )

    if strength_score >= hyperparameters["trend_strength_threshold"]
and trend_metrics["max_streak"] >= 5:
        answer = options[0]
    elif strength_score < 0.4 or trend_metrics["directional_consistency
"] < 0.6:
        answer = options[1]
    elif 0.4 <= strength_score < hyperparameters["
trend_strength_threshold"]:
        answer = options[2]
    else:
        answer = options[3]

    question_text = question.format(ticker=ticker)

    qa_pairs.append({
        "question": question_text,
        "options": options,
        "answer": answer,
        "ticker": ticker,
        "ts": analysis_prices,
        "relevant_concepts": ["Volume-Price Trend Correlation", "Trend
Strength Analysis", "Price Momentum"],
        "domain": "finance",
        "detractor_types": ["Incorrect trend interpretation", "
Misunderstanding momentum signals"],
        "question_type": "multiple_choice",
        "format_hint": "Please answer the question and provide the
correct option letter, e.g., [A], [B], [C], [D], and option content at
the end of your answer. All information need to answer the question is
given. If you are unsure, please provide your best guess.",
    })

return qa_pairs
```

B.3 Example of Natural Language Description

```
I want to create time series exam testing model understanding of finance
    time series data.

To load the data, use the provided ‘‘‘user_dataset‘‘‘ object.

Given time series come from Yahoo Finance, include closing price of a stock.
     Interval between samples is 1 day.
Make sure that the length of time series (total number of samples of one or
     two time series) do not excide 3000.

Please make sure that exams cannot be answer without timeseries!

Time series could include not only raw prices but also derived series such
    as returns, or data from other asset classes beyond stocks.
```

B.4 Examples of Generated Questions

Q: Analyzing the daily price chart of MET (MetLife, Inc.), where in the time series does a significant regime change occur that fundamentally alters the market behavior pattern?

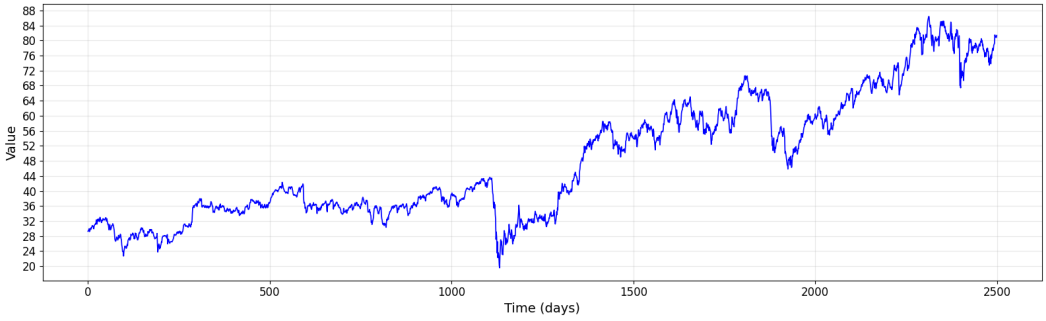
A. Around the beginning of the time series, where the market transitions from one behavioral pattern to another

B. Around the middle of the time series, where there is a clear structural break in volatility or trend patterns

C. Around the end of the time series, where recent market conditions show a distinct change

D. No significant regime change is detectable in this time series

answer: Around the middle of the time series, where there is a clear structural break in volatility or trend patterns



Q: Based on the daily closing price data for MAA over the past 2000 trading days, what does the Relative Strength Index (RSI) analysis reveal about the stock's momentum condition at the end of the period?

A. The stock is in overbought territory with RSI above 70, suggesting potential selling pressure.

B. The stock is in oversold territory with RSI below 30, suggesting potential buying opportunity.

C. The stock shows neutral momentum with RSI around 50, indicating balanced buying and selling pressure.

D. The stock shows strong upward momentum with RSI consistently increasing but not yet overbought.

answer: The stock shows neutral momentum with RSI around 50, indicating balanced buying and selling pressure.

C LLM Verifier

For each template, we use an LLM to evaluate the generated question. Specifically, we ask:

- Is the question relevant to the given concept?
- Does answering the question require the provided time series?
- Are the question and answer free from ambiguity and bias?

C.1 Validation Prompt

```
You are an expert validator of question templates involving reasoning over
{exam_type} time series data.
You are given an exam question template:

{exam_template}

Your task is to validate the question template using the following criteria:
1. Is the question relevant to {exam_type} time series analysis?
2. Would you need the time series itself to answer the question?
3. Are there no ambiguity in the question or its answer?

If the answer to all is YES or MOSTLY YES, return only the number 1.
If the answer to either is NO, return your objections.
Return 1 (do not include any additional text then) or describe your objections.
```

D Other Design Specifics

Detractors In addition, the mechanism of plausible but incorrect answer choices was implemented. The LLM is prompted to reflect on possible mistakes that the test taker might make while solving the exam. Using this knowledge, misleading, incorrect option choices can be generated.

Context Condensation A common issue we encountered in the framework was context window overflow during exam regeneration. To mitigate this, we applied context condensation, which reduces the number of tokens while preserving essential information. In our setup, the agent generates templates in a conversational manner. The process begins with a generation prompt, followed by a message containing the generated exam. If errors occur or the exam is rejected during verification, the feedback and regenerated exams are appended to the conversation. Several context condensation techniques exist, such as windowing [35] and context compression [36]. We adopt a summarization-based method [37, 38], which has shown strong results in prior work and fits our use case. Specifically, we summarize non-recent pairs of failing exams and error messages into short descriptions that highlight the issues encountered. These summaries provide the LLM with concise feedback, supporting the generation of higher-quality templates.

E G-Eval

We evaluated a set of generated questions under the G-Eval framework. We used the following criteria:

```
1. SPECIFICITY
Evaluate the specificity of the generated finance multiple-choice question.

A good finance question should target a single, clearly defined financial
    concept or phenomenon.

Evaluation steps:
1. Read the question and all answer options.
2. Determine if the question targets one specific financial analysis topic
    (e.g., stock trend, valuation ratio, earnings impact).
3. Assess the ratio of finance-specific terms (e.g., "P/E ratio," "support
    level") to generic wording.
4. Penalize if:
- The question is overly broad or open-ended (e.g., "What will happen to
    the market?").
- The wording leaves the intended financial interpretation unclear.
- The question mixes unrelated phenomena (e.g., stock price +
    macroeconomics in one question).

Score highest if the question has one precise focus (e.g., "Did the stock
    break through its resistance level at $150?").

2. UNAMBIGUITY
Evaluate the unambiguity of the generated finance multiple-choice question.

A question and its answers should not allow multiple interpretations.

Evaluation steps:
1. Read the question and all answer options.
2. Determine if the question can be objectively assessed from the financial
     data/context.
3. Check if the answers are clear, distinct, and not overlapping.
4. Penalize if:
- The question uses subjective language (e.g., "Does this stock look risky
    ?").
- The answer options are vague or could mean multiple things.
```

```
- The question cannot be answered with available financial data.

Score highest if the question is clear and objective (e.g., "Did the
    company's EPS increase compared to last quarter?").

3. DOMAIN RELEVANCE
Evaluate the domain relevance of the generated finance multiple-choice
    question.

Does the question pertain to finance, investing, or market analysis?

Evaluation steps:
1. Read the question and all answer options.
2. Identify financial terminology (e.g., "dividend yield," "volatility," "
    earnings report").
3. Determine if the question is relevant to stock/market/financial analysis.

4. Penalize if:
- The question includes non-financial terms or irrelevant topics.
- The question is generic with no financial context.
- The options are not tied to investment/market reasoning.

Score highest if the question contains finance-relevant terms and clearly
    relates to stock price prediction, valuation, or market analysis.

4. ANSWERABILITY
Evaluate the answerability of the generated finance multiple-choice
    question.

Even without the correct answer provided, the question should be answerable
     from available financial data (e.g., price chart, earnings report,
    ratios).

Evaluation steps:
1. Read the question and all answer options.
2. Determine if the question can be answered by analyzing financial
    information (stock price history, fundamentals, reports).
3. Assess whether the question requires actual analysis rather than
    guesswork.
4. Penalize if:
- The question asks about irrelevant factors (e.g., "Was the CEO in a good
    mood?").
- The question can be answered without financial analysis (e.g., trivial
    wording).
- The question is too vague or general.

Score highest if the question requires specific financial analysis (e.g., "
    Did the stock close above its 200-day moving average?").
```

F Hyperparameters

In this section, we list all the hyperparameter used for our agentic workflow.

1. Generator LLM: the LLM use to generate concepts and the corresponding template. We used claude-sonnet-4-20250514 (initial generation with `reasoning_effort="medium"`).
2. Concept LLM: the LLM use to generate concepts. We used gpt-4o-2024-08-06.
3. Verifier LLM: the LLM use to verify templates. We used gpt-4o-2024-08-06.

4. Student LLMs: the student LLMs we use to check the exam differentiability. Currently we have two student LLMs: stronger: gpt-4o-2024-08-06 and weaker: gpt-4o-mini.
5. Exam type: We are generating the data connected to specific domain. We used "ECG" and "finance".
6. Few-shot examples: 3 templates prepared beforehand were used to present the desired structure. For each generation, they were randomly sampled from set of 9.

In such a setup, the generation of one template costs 0.09$ on average.

G Evaluation Protocol

All used models were accessed by API with `LiteLLM` Python library. The following API providers were used with default parameters:

- Closed source models – OpenAI API, Anthropic API
- Open source models – Hugging Face Inference Providers API

During the evaluation, the images of the plots were encoded with base64 encoding and provided to the models. Plots were created with DPI = 50. We used setup without context condensation.