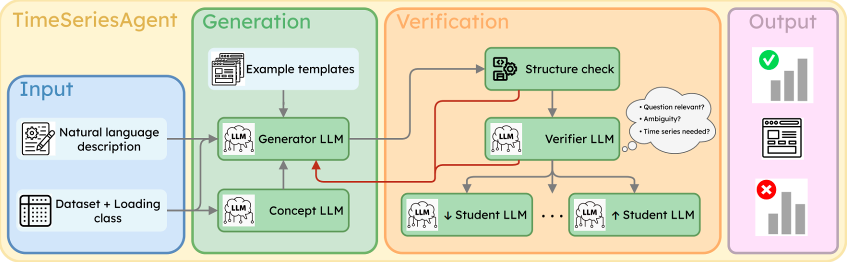
TimeSeriesExamAgent: Creating Time Series Reasoning Benchmarks at Scale

Many recent works have applied Large Language Models (LLMs) to time series analysis tasks such as forecasting, anomaly detection, and classification. At the same time, the topic of evaluating the reasoning capabilities of LLMs in time series tasks have become even more important. However, existing benchmarks have clear limitations. Contextualized tasks remain close to traditional metrics, while reasoning-style benchmarks often focus only on simple properties. In practice, real-world domains such as healthcare require more complex reasoning, where tasks like diagnosis naturally combine anomaly detection, classification, and domain knowledge. Including challenging curation, building specialized, domain-specific benchmarks remains difficult and time-consuming. To address these limitations, we introduce *TimeSeriesExamAgent*, a scalable and domain-agnostic framework for automatically generating, verifying, and refining time series reasoning benchmarks. It enables domain experts to easily create high-quality, domain-specific exams from their own datasets.



The agent takes as input a user's task description in the form of natural language text, dataset and minimal loading code. The agent (architecture shown on the figure) then orchestrates the generation pipeline, including creating question templates, robustness verification from multiple perspectives, and iterative refinement. During the generation stage, LLMs creates question templates (as Python functions) which include time series data from the dataset and multiple-choice questions. Later, templates are automatically checked for structural correctness and verified by an LLM for domain relevance and answerability. If any verification step fails, the template is regenerated with feedback about the error. Finally, capability-aligned filtering is used to discard templates that fail to discriminate between weak and strong answering models. This multistep process allows for generating correct and relevant question templates.

To demonstrate real world usage, we created sets of questions from the domain of healthcare (based on PTB-XL and MIT-BIH ECG datasets) and finance (Yahoo Finance stock dataset). We examined several models, which were chosen to cover a diverse range of performance levels (Gemma-3-27B, GPT-4o, o3-mini, and Qwen2.5-VL-Instruct). We found that while reasoning-oriented models such as o3-mini perform well on finance-related questions, their performance is weaker on healthcare benchmarks, what can suggest that the general reasoning ability does not always transfer across domains.

Empirically, we demonstrate that the framework produces domain-agnostic benchmarks whose diversity matches human generated counterparts. We compare multiple metrics on questions generated from the PTB-XL dataset with those in ECG-QA, a template-based benchmark also built on PTB-XL. Embedding- and edit-distance–based measures show that our framework achieves diversity on par with manually designed datasets. Furthermore, specificity, unambiguity, domain relevance and answerability have been scored during LLM-as-a-judge evaluation, where TimeSeriesExamAgent shows substantially better results.