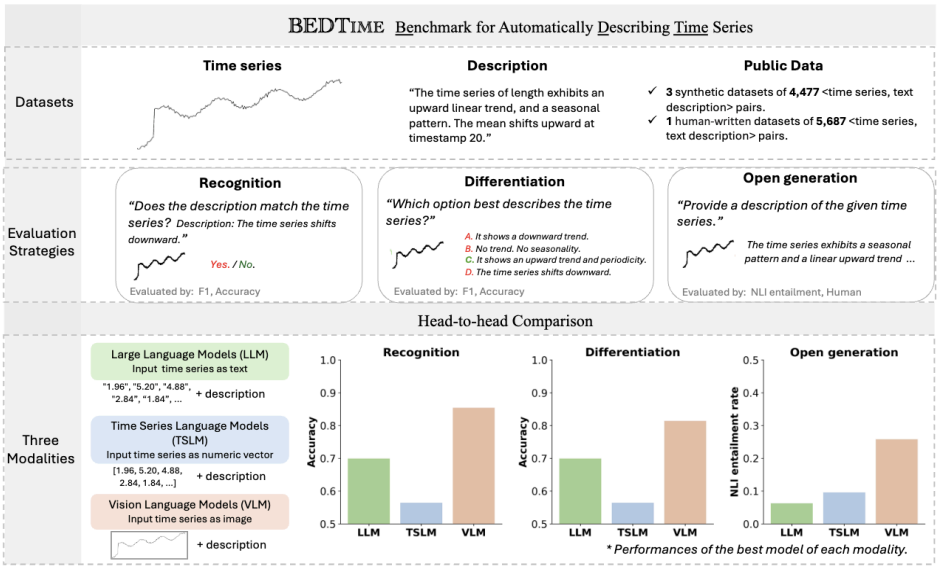
BEDTime: A Unified Benchmark for Automatically Describing Time Series

Time series analysis is central to decision-making in critical domains such as healthcare [1], finance [2], and climate science [3]. Recent foundation models [4-7] promise to interpret time series data, but current evaluations are fragmented: models are introduced alongside specialized datasets, obscuring head-to-head comparisons, and existing benchmarks emphasize complex reasoning tasks rather than fundamental skills. This gap limits our ability to systematically measure progress in multimodal time series reasoning.

To address this gap we introduce **BEDTime**, the **first unified benchmark** designed to test language models'(LM) ability to describe time series in domain-agnostic natural language by decoupling datasets from the models. BEDTime formalizes three fundamental tasks: (1) Recognition (Match a description to a time series), (2) Differentiation (Choose the description that best represents a time series among distractors), and (3) Generation (produce a free-form description). To support these tasks, we unify four diverse univariate datasets — including synthetic and real-world signals — resulting in 10,164 time series–description pairs spanning variable lengths, linguistic styles, and levels of noise. 13 state-of-the-art models are evaluated across three modalities: large language models (LLMs, text input), vision–language models (VLMs, image input), and time series–language models (TSLMs, numeric inputs). Recognition and Differentiation tasks are assessed with accuracy, while Generation is evaluated both by human annotators – using six criteria, all with high inter-rater reliability, and by automatic, robust Natural Language Inference entailment scoring.



Through our evaluation we observe the following: First, LLMs underperform on descriptive tasks, highlighting limits of text-only representations. Second, VLMs are surprisingly strong, often achieving near-perfect accuracy on synthetic datasets, but still fall short on real-world data. Third, TSLMs can rival or surpass open-source LLMs of comparable parameter size, especially on long and structured sequences, but remain weaker than VLMs overall. Importantly, all models break down under robustness tests—including Interpolation, Interpolation-Scaling, Amplitude Scaling, Missing Data, and Additive Gaussian Noise—revealing consistent fragility to longer sequences, missing values, and real-world perturbations. However, for LLMs, chain-of-thought (CoT) prompting provided noticeable gains, partially mitigating some weaknesses.

Overall, BEDTime provides the first rigorous and systematic framework for evaluating the time series descriptive ability of LMs. It establishes a standardized platform for head-to-head comparison across three modalities. Our findings show the promise of time-series–aware models and the persistent fragility of all approaches under realistic perturbations, thereby offering guidance for the next generation of multimodal architectures. In addition, we release a flexible code base that enables researchers to integrate new models or adapt novel time series–text datasets into the BEDTime framework, supporting future extensibility and development.

[1] Tomov et al. (2023). Time series in epidemiology: Review and COVID-19 experience. World J Clin Cases, 11(29), 6974–6983.

[2]Fang et al. (2025). Deep neural network modeling for financial time series analysis. Big Data Research, 41, 100553.

[3]Volvach et al. (2022). Time series analysis of temperatures and insolation of the Earth's surface at Kara-Dag using satellite observation. Advances in Space Research, 69(12), 4228–4239.

[4] Wang et al. (2025). ChatTime: A unified multimodal time series foundation model. AAAI.

[5] Xie et al. (2024). ChatTS: Aligning time series with LLMs via synthetic data. arXiv:2412.03104.

[6] Trabelsi et al. (2025). Time series language model for descriptive caption generation. arXiv:2501.01832

[7] Chow et al. (2024). Towards time-series reasoning with LLMs. NeurIPS.