Towards Interpretable Time Series Foundation Models

Matthieu Boileau [1] **Philippe Helluy** [1] **Jérémy Pawlus** [2] **Svitlana Vyetrenko** [3 1]

Abstract

In this paper, we investigate the distillation of time series reasoning capabilities into small, instruction-tuned language models as a step toward building interpretable time series foundation models. Leveraging a synthetic dataset of mean-reverting time series with systematically varied trends and noise levels, we generate natural language annotations using a large multimodal model and use these to supervise the fine-tuning of compact `Qwen` models. We introduce evaluation metrics that assess the quality of the distilled reasoning—focusing on trend direction, noise intensity, and extremum localization—and show that the post-trained models acquire meaningful interpretive capabilities. Our results highlight the feasibility of compressing time series understanding into lightweight, language-capable models suitable for on-device or privacy-sensitive deployment. This work contributes a concrete foundation toward developing small, interpretable models that explain temporal patterns in natural language.

1. Introduction

1.1. Problem statement

Recent research has shown that small models can achieve performance comparable to much larger models when trained on carefully curated datasets. For instance, (Gunasekar et al., 2023) demonstrate that high-quality, instruction-tuned data can dramatically improve model capabilities without scaling up parameters. Similarly, (Eldan & Li, 2023) highlights that small language models trained on compact, well-structured data can generate coherent and creative narratives. This principle is further reinforced by advancements like `SKY T1` (Team, 2025) and `DeepSeek R1` (DeepSeek-AI et al., 2025), where focused data selection and training techniques enable relatively lightweight models to perform tasks traditionally reserved for larger architectures. Together, these studies suggest that data quality, not just model size, is a critical factor in achieving strong language model performance.

Time series data is ubiquitous across healthcare, finance, and industrial systems, where real-time, on-device analysis is essential for operational efficiency and timely decision-making. Models in these settings must address a range of tasks, including forecasting, anomaly detection, trend analysis, and event classification. Beyond performance and compactness, it is increasingly important for such models to be interpretable—able to explain temporal patterns and anomalies in natural language that aligns with human reasoning. Despite its significance, this direction remains largely unexplored. In this work, we take a step toward this goal by introducing a method for distilling time series reasoning into small, interpretable models capable of generating structured natural language explanations.

1.2. Related work

Recent work has investigated the application of general-purpose language models to time series tasks. Gruver and Wilson (Gruver et al., 2024) first demonstrated that pre-trained language models can achieve competitive zero-shot performance on standard forecasting benchmarks. Building on this, Jin et al. (Jin et al., 2024) proposed a broader framework in which a variety of time series tasks are reformulated as token prediction problems within a language modeling paradigm. In parallel, specialized foundational models pre-trained exclusively on time series data have been introduced. Chronos (Ansari et al.), Llaglama (Rasul et al., 2024), and TimesFM (Das et al., 2024) focus on large-scale pretraining for forecasting, achieving strong zero-shot and few-shot performance across diverse temporal datasets. Moirai (Woo et al., 2024) extends these methods to multivariate forecasting, capturing complex dependencies among multiple correlated series. Multitask models such as UniTS (Gao et al., 2024) and Moment (Goswami et al., 2024) further generalize this approach by jointly training across forecasting, imputation, and classification tasks, illustrating the advantages of shared temporal representations.

[1]Univeristy of Strasbourg, France [2]AxesSim, Strasbourg, France [3]Outsampler, Strasbourg, France. For correspondence and code requests email: Svitlana Vyetrenko <svitlana@outsampler.com>.

International Conference on Machine Learning (ICML) 2025 Workshop on Foundation Models for Structured Data, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

Systematic evaluations based on the series feature taxonomy proposed in (Fons et al., 2024b) benchmark general-purpose language models on a range of time series understanding tasks, highlighting both their strengths and limitations relative to specialized models. TimeSeriesExam (Cai et al., 2024) introduces a complementary benchmark focused on evaluating language models' ability to perform detailed reasoning over temporal patterns. In parallel, work in (Daswani et al., 2024) extends language model-based methods to the interpretation of visual temporal data, including line and bar charts. Together, these efforts reflect a broader shift toward treating time series and structured data as instances of language modeling.

1.3. Our contributions

1. We propose a framework for **interpretable time series foundational model** construction via knowledge distillation, enabling small language models to reason about time series data in natural language.

2. We introduce a **practical and reproducible distillation pipeline** that transfers simple time series reasoning skills—such as trend identification, noise assessment, and extrema localization —- from large multimodal models to compact ones.

3. We contribute a **baseline dataset and evaluation framework** for assessing time series reasoning in small models, providing a foundation for future work on interpretable time series foundation models.

2. Distillation of time series reasoning

2.1. Background

Distillation is the process of transferring knowledge from a large, often complex model (the teacher) to a smaller, more efficient model (the student) that can understand and reason about a narrow topic (Xu et al., 2024; Hinton et al., 2015). This technique aims to retain the performance of the larger model while reducing computational requirements, making deployment more practical in real-time or resource-constrained settings. In addition, practical benefits can include the ability to fine-tune smaller models with sensitive data.

Recent advances in distilling reasoning—particularly mathematical inference—into small models have demonstrated that compact architectures can tackle complex tasks with surprising accuracy(Team, 2025; DeepSeek-AI et al., 2025). Distillation of time series reasoning has not been explored yet in the literature. In this paper, we explore this capability and demonstrate that small models can be effectively trained via distillation.

2.2. Synthetic dataset generation and annotation

We create an annotated dataset of time series given by Orstein-Uhlenbeck process (Byrd, 2019):

$$r_t = r_{t-1} + \kappa(\overline{r} - r_{t-1}) + u_t, \quad r_0 = 0, \tag{1}$$

where $\overline{r}$ is a mean value of the process, κ is a mean-reversion parameter and $u_t \sim \mathcal{N}(0, \sigma^2)$ is random noise added to the time series at each time step t.

Orstein-Uhlenbeck process is frequently used to model financial markets (Wah et al., 2017; Bamford et al., 2024) as well as biological processes (Oksendal, 1998). As argued in (Zhou & Yu, 2025), large language models generally perform better in time series reasoning when processing time series data as images rather than text tokens; therefore, we use both numerical and visual time series presentations in order to generate high-quality time series annotations by a large model. We generate time series captions by prompting a large model `pixtral-large` by Mistral AI to generate three sentences that describe the time series trend (further referred to as "trend"), possible presence and intensity of noise (further referred to as "noise") and description of local and global extrema (further referred to as "extrema"). See Figures 1, 2 and 3 for the sample time series image input, prompt and the generated output. Using the above procedure, and varying randomly the parameters $(\kappa, \overline{r}, \sigma)$ we generate 200 time series-image-annotation samples. Then we will allocate 180 samples for training, and 20 samples for testing.

2.3. Evaluation metrics

In order to evaluate the quality of the dataset and, subsequently, the quality of the post-trained small models, we propose the following evaluation metrics:

1. **Standard cosine similarity metric** between the entire annotated paragraphs (that is, the three sentences) computed with the sentence-transformers model `all-MiniLM-L6-v2`. Cosine similarity score is known to measure overall relatedness and syntactic similarity in the annotation style, however, might not be able to capture a strong contradiction in the logical meaning (e.g., "Time series is increasing" and "Time series is decreasing").

2. **NLI scores** on a per-sentence level computed with `roberta-large-mnli` model. Unlike cosine similarity score above, NLI scores help highlight if there is a strong contradiction between the generated annotations and the ground truth (Williams et al., 2018). For each of the three diagnostic sentences ("trend", "noise", "extrema"), we attribute 0 to contradiction, 0.5 to neutral and 1 to entailment and then compute the average score across the samples.

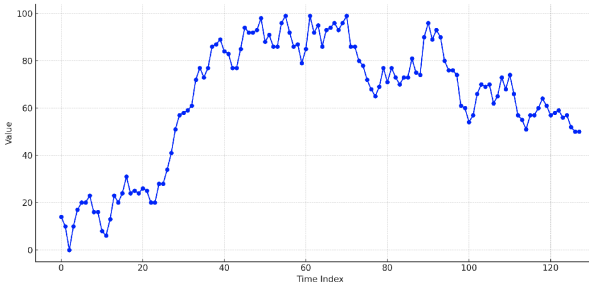

Figure 1: Sample time series image input to `pixtral-large`.

PROMPT: Describe the time series in three sentences. First sentence: describe trend (increasing/decreasing/flat). Second sentence: noise intensity (low/medium/high). Third sentence: approximate localisation of global maximum (beginning/middle/end) and global minimum (beginning/middle/end). Put the description in a JSON format with the following pattern { "trend": sentence1 , "noise": sentence2 , "extrema": sentence3 }

Figure 2: Prompt used to annotate time series by `pixtral-large`. It is worth noting that such prompting is requiring the model to simply localize the extrema, without explicitly specifying them.

ANNOTATION: The time series exhibits an increasing trend initially, followed by fluctuations and a general decreasing trend towards the end. The noise intensity in this time series is high, with significant fluctuations throughout. The global maximum is approximately located in the middle of the time series, while the global minimum is towards the beginning.

Figure 3: Sample time series annotation generated by `pixtral-large`.

3. **Feature-based similarity scores** We assess the quality of annotations by comparing them to fact-based reference sentences computed based on explicit derivation of time series features using the following procedure. First, the trend is determined based on the overall increasing or decreasing behavior of a smoothed version of the time series. Second, the noise level is estimated from the variance of the data. Third, the positions of global extrema are computed precisely in order to describe their approximate locations. These three analyses are used to generate fact-checked sentences such as: "The time series shows an overall increasing/flat/decreasing trend"; "The noise intensity is low/medium/high"; "The minimum/maximum occurs around the beginning/middle/end of the time series".

2.4. Fact-checking

To improve the confidence in our dataset, we perform a NLI comparison between `pixtral-large` and our feature-based annotations. For the "trend" and "noise" annotations, we observe agreement rates of 93% and 95%, respectively, between `pixtral-large` and the fact-based sentences. The few detected discrepancies generally reflect acceptable ambiguity. For example, a time series with large fluctuations might reasonably be described as either "increasing" or exhibiting "no clear trend." We choose to retain such cases in the dataset. No unambiguous contradictions (such as simultaneous claims of "increasing" and "decreasing" trends) were observed, which aligns with the documented performance of `pixtral-large` on the ChartQA benchmark (pix).

For the "extrema" annotation, the NLI comparison identifies 5% of cases as contradictions. Some of these reflect genuine inconsistencies. In such cases, we replace the annotation with the corresponding fact-based sentence to ensure the overall quality of the dataset.

2.5. Small model training

We use the curated dataset to fine-tune `Qwen2.5-1.5B-Instruct` and `Qwen2.5-0.5B-Instruct` models (Qwen et al., 2025). Such models are equipped with basic natural language capabilities, however, due to their size they are expected to perform worse on time series reasoning without post-training (as our subsequent testing then confirms - see Table 1). At this stage, we fine-tune the small `Qwen` models using only the numerical time series data paired with their annotated text, and without including time series images.

It should be noted that tokenization of numerical values poses challenges for large language models (LLMs), as most tokenizers are optimized for natural language and not for numeric precision (Spathis & Kawsar, 2023). Decimal numbers, scientific notation, and long integers are often split into multiple tokens inconsistently, which can hinder model performance on tasks involving quantitative reasoning. To ensure consistent tokenization, we rescale and round all floating numbers obtained by the time series generation process to integers with left padding from 00 to 99, and ensure that each digit is represented as a single token (Yuan et al., 2023).

2.6. Experimental results and evaluation

To evaluate the quality of time series reasoning that we transfer to small models `Qwen2.5-1.5B-Instruct` and `Qwen2.5-0.5B-Instruct` through the distilled samples, we compare the annotations they generate before and after post-training with the initial annotations generated by

	Qwen2.5-1.5B	Qwen2.5-0.5B
Cosine	0.5	0.62
NLI "trend"	0.7	0.6
NLI "noise"	0.575	0.275
NLI "extrema"	0.35	0.475
Feature "trend"	0.8	0.6
Feature "noise"	0.475	0.45
Feature "extrema"	0.475	0.45

Table 1: Evaluation results before post-training. Note that prior to training, the small models are unable to produce valid JSON outputs in the expected three-sentence format. Therefore, for the pre-training evaluation, we assess their performance by asking separate questions about trend, noise, and extrema, without requiring JSON formatting.

	Qwen2.5-1.5B	Qwen2.5-0.5B
Cosine	0.98	0.98
NLI "trend"	0.875	0.875
NLI "noise"	0.75	0.675
NLI "extrema"	0.875	0.775
Feature "trend"	1.0	1.0
Feature "noise"	0.8	0.8
Feature "extrema"	0.95	0.95

Table 2: Evaluation results after post-training

`pixtral-large` on the test dataset using the evaluation metrics described in Section 2.3.

As evident from Table 1, we observe that initially the small `Qwen` models do not display much time series reasoning. Notably, small models understand trends better than other features which agrees with test results for large models in (Fons et al., 2024a) and (Cai et al., 2024). The experimental results were significantly improved after post-training as one can observe in Table 2. Therefore, our experiments suggest that even very small models are capable of learning basic time series reasoning in natural language.

3. Conclusion

Our findings demonstrate that distillation of time series reasoning is feasible even for very small models, enabling effective downstream use in resource-constrained environments such as wearable devices. Such post-trained models can be further fine-tuned with application-specific data, and they exhibit promising utility in real-world scenarios.

Our future work will focus on enhancing interpretable foundation model for time series capabilities through visual and numerical representation learning, exploring alternative numerical tokenization strategies for effective floating-point representation, and also improving precision in arithmetic reasoning and retrieval. Additionally, small post-trained models equipped with time-series specific reasoning capability offer a useful testbed for observing and understanding the out-of-training-distribution generalization, which we plan to explore in greater depth in our future work.

Acknowledgements

This work has been supported by the Institut Thematique Interdisciplinaire IRMIA++ at the University of Strasbourg (https://irmiapp.unistra.fr/) and the Gutenberg Circle.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Available from https://mistral.ai/news/pixtral-large.

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., Zschiegner, J., Maddix, D. C., Wang, H., Mahoney, M. W., Torkkola, K., Wilson, A. G., Bohlke-Schneider, M., and Wang, Y. Chronos: Learning the language of time series.

Bamford, T., Coletta, A., Fons, E., Gopalakrishnan, S., Vyetrenko, S., Balch, T., and Veloso, M. Multi-modal financial time-series retrieval through latent space projections, 2024. URL https://arxiv.org/abs/2309.16741.

Byrd, D. Explaining agent-based financial market simulation, 2019. URL https://arxiv.org/abs/1909.11650.

Cai, Y., Choudhry, A., Goswami, M., and Dubrawski, A. Timeseriesexam: A time series understanding exam, 2024. URL https://arxiv.org/abs/2410.14752.

Das, A., Kong, W., Sen, R., and Zhou, Y. A decoder-only foundation model for time-series forecasting, 2024. URL https://arxiv.org/abs/2310.10688.

Daswani, M., Bellaiche, M. M. J., Wilson, M., Ivanov, D., Papkov, M., Schnider, E., Tang, J., Lamerigts, K., Botea, G., Sanchez, M. A., Patel, Y., Prabhakara, S., Shetty, S., and Telang, U. Plots unlock time-series understanding

in multimodal models, 2024. URL https://arxiv.org/abs/2410.02637.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Ye, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Xu, Y., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Eldan, R. and Li, Y. Tinystories: How small can language models be and still speak coherent english?, 2023. URL https://arxiv.org/abs/2305.07759.

Fons, E., Kaur, R., Palande, S., Zeng, Z., Balch, T., Veloso, M., and Vyetrenko, S. Evaluating large language models on time series feature understanding: A comprehensive taxonomy and benchmark. *arXiv preprint arXiv:2404.16563*, 2024a.

Fons, E., Kaur, R., Palande, S., Zeng, Z., Balch, T., Veloso, M., and Vyetrenko, S. Evaluating large language models on time series feature understanding: A comprehensive taxonomy and benchmark, 2024b. URL https://arxiv.org/abs/2404.16563.

Gao, S., Koker, T., Queen, O., Hartvigsen, T., Tsiligkaridis, T., and Zitnik, M. Units: A unified multi-task time series model, 2024. URL https://arxiv.org/abs/2403.00131.

Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., and Dubrawski, A. Moment: A family of open time-series foundation models, 2024. URL https://arxiv.org/abs/2402.03885.

Gruver, N., Finzi, M., Qiu, S., and Wilson, A. G. Large language models are zero-shot time series forecasters, 2024. URL https://arxiv.org/abs/2310.07820.

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Giorno, A. D., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Behl, H. S., Wang, X., Bubeck, S., Eldan, R., Kalai, A. T., Lee, Y. T., and Li, Y. Textbooks are all you need, 2023. URL https://arxiv.org/abs/2306.11644.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Jin, M., Zhang, Y., Chen, W., Zhang, K., Liang, Y., Yang, B., Wang, J., Pan, S., and Wen, Q. Position: What can large language models tell us about time series analysis, 2024. URL https://arxiv.org/abs/2402.02713.

Oksendal, B. Stochastic differential equations, , an introduction with applications. 1998.

Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Rasul, K., Ashok, A., Williams, A. R., Ghonia, H., Bhagwatkar, R., Khorasani, A., Bayazi, M. J. D., Adamopoulos, G., Riachi, R., Hassen, N., Biloš, M., Garg, S., Schneider, A., Chapados, N., Drouin, A., Zantedeschi, V., Nevmyvaka, Y., and Rish, I. Lag-llama: Towards foundation models for probabilistic time series forecasting, 2024. URL https://arxiv.org/abs/2310.08278.

Spathis, D. and Kawsar, F. The first step is the hardest: Pitfalls of representing and tokenizing temporal data for large language models, 2023. URL https://arxiv.org/abs/2309.06236.

Team, N. Sky-t1: Train your own o1 preview model within 450. $https://novasky-ai.github.io/posts/sky-t1, 2025. Accessed: 2025-01-09$.

Wah, E., Wright, M., and Wellman, M. P. Welfare effects of market making in continuous double auctions. 2017.

Williams, A., Nangia, N., and Bowman, S. R. A broad-coverage challenge corpus for sentence understanding through inference, 2018. URL https://arxiv.org/abs/1704.05426.

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., and Sahoo, D. Unified training of universal time series forecasting transformers, 2024. URL https://arxiv.org/abs/2402.02592.

Xu, X., Li, M., Tao, C., Shen, T., Cheng, R., Li, J., Xu, C., Tao, D., and Zhou, T. A survey on knowledge distillation of large language models, 2024. URL https://arxiv.org/abs/2402.13116.

Yuan, Z., Yuan, H., Tan, C., Wang, W., and Huang, S. How well do large language models perform in arithmetic tasks?, 2023. URL https://arxiv.org/abs/2304.02015.

Zhou, Z. and Yu, R. Can llms understand time series anomalies?, 2025. URL https://arxiv.org/abs/2410.05440.