A Theoretical Analysis of Detecting Large Model-Generated Time Series

Junji Hou, Junzhou Zhao*, Shuo Zhang, Pinghui Wang

MoE KLINNS Lab, Xi'an Jiaotong University, Xi'an 710049, P. R. China
{15955192, zs412082986}@stu.xjtu.edu.cn, {junzhou.zhao, phwang}@xjtu.edu.cn

Abstract

Motivated by the increasing risks of data misuse and fabrication, we investigate the problem of identifying synthetic time series generated by Time-Series Large Models (TSLMs) in this work. While there is extensive research on detecting model generated text, we find that these existing methods are not applicable to time series data due to the fundamental modality difference, as time series usually have lower information density and smoother probability distributions than text data, which limit the discriminative power of token-based detectors. To address this issue, we examine the subtle distributional differences between real and model-generated time series and propose contraction hypothesis, which states that model-generated time series, unlike real ones, exhibit progressively decreasing uncertainty under recursive forecasting. We formally prove this hypothesis under theoretical assumptions on model behavior and time series structure. Model-generated time series exhibit progressively concentrated distributions under recursive forecasting, leading to uncertainty contraction. We provide empirical validation of the hypothesis across diverse datasets. Building on this insight, we introduce the Uncertainty Contraction Estimator (UCE), a white-box detector that aggregates uncertainty metrics over successive prefixes to identify TSLM-generated time series. Extensive experiments on 32 datasets show that UCE consistently outperforms state-of-the-art baselines, offering a reliable and generalizable solution for detecting model-generated time series.

1 Introduction

Recent advances in time series forecasting have given rise to Time Series Large Models (TSLMs), which are pre-trained on massive multi-domain time series datasets with billions of parameters (Ansari et al. 2024; Liu et al. 2024; Shi et al. 2025). The vast training data and enormous parameter scale enable TSLMs to achieve remarkable long-term zero-shot forecasting on previously unseen datasets or domains without any labeled examples or task-specific fine-tuning, as evidenced by recent scaling law analyses (Yao et al. 2025). Leveraging these sophisticated forecasting capabilities, TSLMs have demonstrated strong performance in domains such as finance, the Internet of Things (IoT), and climate science.

The powerful capability of TSLMs raises significant concerns about potential data fabrication or misuse. Unlike classical approaches (e.g., ARIMA, LSTM), which degrade in performance outside their training domains, TSLMs can generate coherent long sequences even for unfamiliar domains. If maliciously exploited, this capability could enable the systematic fabrication of time series and pose severe threats in scenarios where data authenticity is critical.

- **Finance:** TSLMs can synthesize long transaction histories that mirror real trading patterns, facilitating fraudulent activities such as inflated valuations or hidden manipulations, including the 2012 LIBOR scandal (Gupta 2024; Rose and Sesia 2013).
- **Scientific Research:** Highly realistic counterfeit measurement time series (e.g., signal traces or biological data) can distort experimental outcomes, similar to Schön and Wakefield data forgeries (Brumfiel 2002; Godlee, Smith, and Marcovitch 2011).
- **Environmental Governance:** Attackers can generate counterfeit metrics (e.g., air quality indices or emissions) that reproduce genuine diurnal and seasonal cycles to conceal pollution spikes or overstate improvements, misleading policymakers and obscuring real hazards (Wang, Wang, and Wilkes 2021).

To address these threats, we introduce a theoretical framework for white-box detection of TSLM-generated time series. Since there are no dedicated detection methods for time series data, we adapt text-based detection methods to time series. Textual detectors exploit the observation that LLMs exhibit different token-level probability patterns on human-written and model-generated text, and therefore typically build zero-shot classifiers from each token's probability or its rank within the model's vocabulary. However, these methods face significant challenges when applied to time series because of fundamental modality differences.

Textual data exhibit structural semantics and carry rich token-level information (Meister et al. 2022), which create greater semantic distances among different tokens. In any given context, only a small subset of tokens is semantically plausible and has higher probability (e.g., probabilities might concentrate on "apple", "orange" given the prefix

*Corresponding Author.

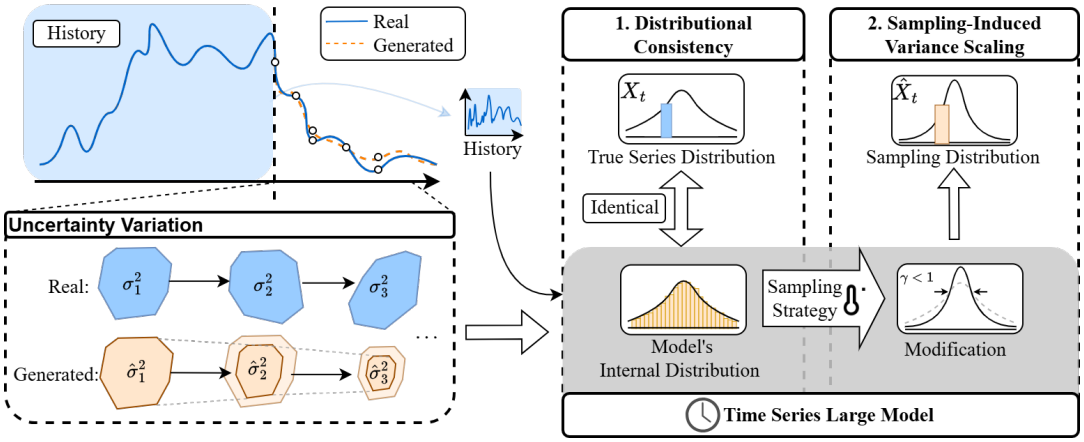


Figure 1: Illustration of the variation in uncertainty for real and model-generated time series.

"I eat an"). This produces sharper, low-entropy probability distributions over the vocabulary, with a narrower range of token selection.

In contrast, time series carry less information at each point (Nie et al. 2023) and inherently contain greater intrinsic uncertainty, yielding smoother probability distributions. Notably, despite the greater entropy of these smooth distributions, they reflect the information at each time point rather than specific values. Since the adjacent values in time series are highly similar (e.g. temperature $25.1°\mathrm{C}$ and $25.2°\mathrm{C}$), which results in large mutual information, time series values convey less information. Consequently, text-based detectors that rely on token probabilities perform poorly on time series given the relatively lower probability gaps between values.

Since point-wise probabilities are insufficiently discriminative in time series, we instead analyze the full distributions from the model. We show that a TSLM's internal distributions, conditioned on the full history, accurately capture the true series distributions and their inherent uncertainty. As TSLMs are trained to minimize prediction errors, the internal distributions are concentrated via the model's sampling strategies for forecasting. In recursive forecasting, uncertainty is cumulatively decreasing, leading to progressively concentrated internal distributions, as illustrated in Fig. 1.

We therefore propose the **contraction hypothesis**, i.e., TSLM-generated time series exhibit progressively decreasing uncertainty, whereas real time series do not. To validate this hypothesis, we provide a theoretical analysis under idealized assumptions on model behavior and time series structure (see Section 4.2) with empirical evidence through long-horizon forecasting experiments (see Fig. 2). Grounded in this detailed analysis of properties unique to time series, we introduce the Uncertainty Contraction Estimator (UCE), a white-box model generation detection method for time series data. UCE captures uncertainty dynamics from internal prediction distributions and identifies sequences with lower uncertainty levels as model-generated.

Our main contributions are summarized as follows.

- To the best of our knowledge, we present the first framework for white-box detection of TSLM-generated time series. We analyze the detailed properties unique to time series in contrast to textual data and address the challenge of low information density in time series.
- We propose the contraction hypothesis, which states that model-generated time series exhibit progressively decreasing uncertainty during recursive forecasting, whereas real series do not. We provide a theoretical analysis under idealized assumptions of time series and model and empirically validate it.
- Based on this hypothesis, we develop the Uncertainty Contraction Estimator (UCE), which captures uncertainty dynamics over successive prefixes using TSLMs to distinguish model-generated from real time series.

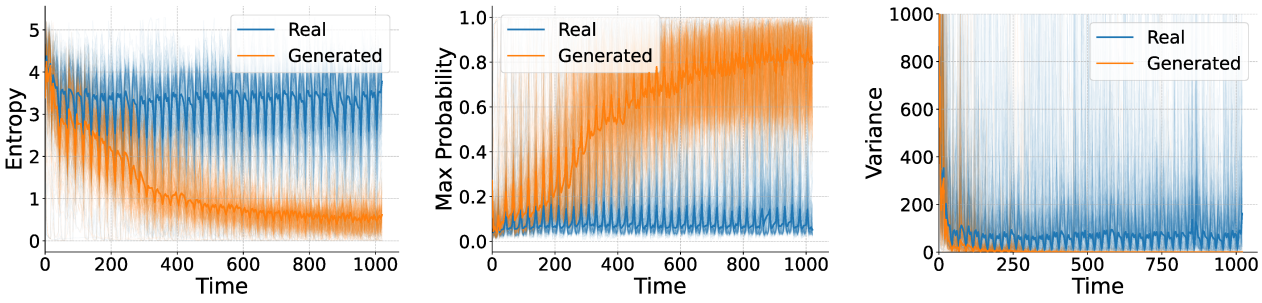
2 Related Work

2.1 Time Series Large Models

Time series large models have emerged to significantly advance zero-shot time series forecasting. Earlier methods such as PromptCast (Xue and Salim 2023) and LLM-Time (Nate et al. 2023) directly leverage LLMs for time series. They convert time series data into text-based prompts and use pretrained LLMs with little to no task-specific adaptation. Such methods require dataset-specific templates and rely heavily on model scale.

PatchTST (Nie et al. 2023) introduces patchifying time series into embeddings, a technique later adopted in models such as GPT4TS (Zhou et al. 2023) and Time-LLM (Jin et al. 2024). These methods use patch embeddings as inputs to LLMs for prediction, but require extensive fine-tuning. Moreover, their performance and inference efficiency have been questioned (Tan et al. 2024).

Recent work concentrates on models trained on vast and diverse time series datasets. MOMENT (Goswami et al. 2024) employs uniform random masking to generate patch embeddings for self-supervised pretraining. Chronos (Ansari et al. 2024) discretizes real-valued time series through scaling and quantization for forecasting and introduces data augmentation to improve generalization. Timer (Liu et al. 2024) adopts a decoder-only transformer



(a) Entropy (b) Max-Probability (c) Variance

Figure 2: The empirical results show the trajectories of uncertainty metrics, including entropy (2a), max-probability (2b) and variance (2c) of both real and model-generated time series data, illustrating reduction in uncertainty for generated data.

architecture, enabling multiple temporal tasks such as forecasting, anomaly detection, and imputation. Time-MoE (Shi et al. 2025) leverages a Mixture-of-Experts framework to improve forecasting performance and enhance cross-domain generalization while maintaining computational efficiency. These methods collectively represent the emerging trend of general-purpose time series models capable of handling diverse tasks with minimal or no fine-tuning.

2.2 Model Generation Detection

The growing generative capabilities of LLMs have raised concerns about distinguishing model-generated content, particularly in the textual domain. Early work on model generation detection conducts supervised classification using bag of words (Solaiman et al. 2019) or neural representations (Jawahar, Abdul-Mageed, and Lakshmanan 2020; Uchendu et al. 2020), but often overfit and underperform on out-of-distribution data (Pu et al. 2023). To overcome this limitation, zero-shot detectors use LLM output statistics, such as perplexity (Lavergne, Urvoy, and Yvon 2008) and log rank (Gehrmann, Strobelt, and Rush 2019; Hashimoto, Zhang, and Liang 2019).

Recent work analyzes token probabilities and uses the discrepancies between human-written and model-generated text to develop classifiers. DetectGPT (Mitchell et al. 2023), Fast-DetectGPT (Bao et al. 2023), and NPR (Su et al. 2023) compare probability differences between perturbed texts, while DNA-GPT (Yang et al. 2024) leverages regeneration to compute divergence. FourierGPT (Xu et al. 2024) performs a spectral analysis on token probability sequences to extract linguistic features that distinguish human-authored from model-generated text. Binocular (Hans et al. 2024) leverages both an observer and a performer to compute cross-perplexity and isolate intrinsic prefix-induced uncertainty to improve detection accuracy. Black-box methods such as intrinsic dimension (Tulchinskii et al. 2023) perform topological data analysis over token embeddings and find that human-written texts are more complicated in topology.

Despite their success in text, these methods rely on the dense and discrete structure of language, making them unsuitable for continuous and information-sparse time series data. In this work, we extend zero-shot model generation detection to time series by introducing uncertainty-based metrics derived from model's internal probability distributions.

3 Problem Formulation and Preliminaries

In this section, we formally define the model-generated time series detection problem. Then we provide some preliminaries about our framework.

3.1 Problem Formulation

Let $\mathbf{X}_t = (X_1, \ldots, X_t)$ denote a univariate time series with *unknown* history. We perform a zero-shot classification of $\mathbf{X}_t$ as real or model-generated without labeled examples.

We assume a white-box setting with access to the model's internal probability distribution $p_\theta(\cdot|\mathbf{X}_t)$ rather than the model architecture or parameters θ. Our detection employs point-wise probabilistic TSLMs over an equidistant vocabulary $\mathcal{V} = \{v_i | v_{i+1} - v_i = \Delta, -R \leq v_i \leq R\}$ for new token sampling. Therefore, sampling token v_i yields the numerical forecast $\hat{X}_{t+1} = v_i$. In the limit $\Delta \to 0$ and $R \to \infty$, $\mathcal{V}$ converges to $\mathbb{R}$ and induces a probability density f_θ. Such a formulation allows for detection evaluation across diverse model architectures.

Some frequently used symbols throughout the paper are given in Table 1. In the following sections, we refer to distributions by their probability densities.

3.2 Preliminaries

We first introduce some assumptions and premises about time series and idealized TSLMs. Detailed definitions and assumptions are provided in the Appendix.

We formally model time series as realizations of stochastic processes, where each observation sequence is a sample path of this process. We decompose the real time series process at time point t into two components (Box et al. 1978):

$$X_t = T_t + n_t, \quad n_t \sim \mathcal{N}\left(0, \sigma_t^2\right),$$

where T_t denotes the trend sequence, representing predictable components, and n_t is a Gaussian process with zero mean and variance such that $\sigma_t^2 = \sum_{i=1}^{l} \alpha_i \sigma_{t-i}^2$, $\sum_{i=1}^{l} \alpha_i = 1$. The noise process represents the unpredictable or uncertainty components.

Symbol	Description
$X_t, \hat{X}_t$	Real / forecast time series process
T_t	Deterministic trend component at time t
$n_t, \hat{n}_t$	True / forecast (Gaussian) noise at time t
$\mathbf{X}_{-H:t}, \hat{\mathbf{X}}_{-H:t}$	History $X_{-H}, \ldots, X_t$ (or $\hat{X}_{-H}, \ldots, \hat{X}_t$)
$f_t(X_t), \hat{f}_\theta(\hat{X}_t)$	Probability density of X_t / $\hat{X}_t$ at time t
$\sigma_t^2, \hat{\sigma}_t^2$	Variance of true noise n_t / forecast noise $\hat{n}_t$
$f_\theta(z_t \mid \mathbf{X}_{-H:t-1})$	Model's internal probability density with history $\mathbf{X}_{-H:t-1}$, abbreviated as $f_\theta(z_t)$
$\tilde{\sigma}_t^2$	Variance of internal probability distribution at time t
γ_t	Scaling factor of sampling strategy to $f_\theta(z_t)$ at time t

Table 1: Some frequently used symbols

We introduce *Ideal Model*, which, for any real-valued time series in $\mathbb{R}$, predicts probability distributions to minimize the expectation of cross-entropy loss. Conceptually, it generalizes practical TSLMs by letting both training data and model capacity grow without bound. This abstraction allows us to derive fundamental detection principles regardless of specific architectures or training constraints.

We formalize *model evaluation function* as Eq. (1):

$$\text{Eva} = \left[\sum_{i=1}^{\tau} g\left(\left|X_i - \hat{X}_i\right|^p\right)\right]^{1/p}, 1 \le p \le \infty, \quad (1)$$

where g is any non-negative strictly increasing mapping. Specifically, if g is linear, Eva reduces to the standard ℓ_p norm of point-wise errors (e.g., RMSE when $p = 2$). Since X_i and $\hat{X}_i$ are variables, Eva is also a variable. Eva is a unified form of various model prediction evaluators (e.g., MSE, MAE) which in turn guides the model generation process.

4 Methodology

In this section, we introduce the Uncertainty Contraction Estimator (UCE), a novel time series model generation detection method. We first establish the theoretical foundation by proposing the *contraction hypothesis* in Section 4.1, followed by a formal analysis in Section 4.2. Finally, the implementation details of UCE are presented in Section 4.3.

4.1 Contraction Hypothesis

UCE hinges on two observations: (1) the model's internal distributions faithfully reproduce true series distributions at each time; (2) through sampling strategies, the model systematically modifies these distributions, creating a self-reinforcing process of exponentially decreasing uncertainty. We formalize these insights as the contraction hypothesis.

Contraction hypothesis: TSLM-generated time series exhibit progressively concentrating internal distributions with decreasing uncertainty, whereas real series do not.

The term "contraction" refers to the contraction mapping nature of the uncertainty reduction. We systematically substantiate the hypothesis through a tripartite theoretical framework. Formal proofs of all the following propositions are provided in the Appendix.

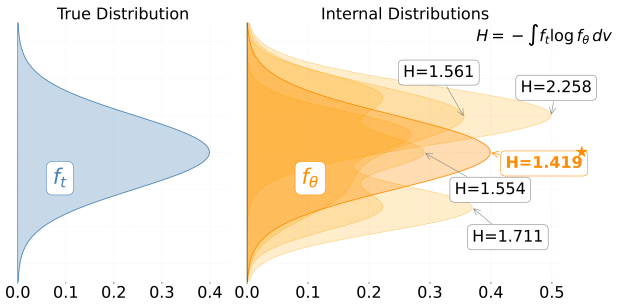


Figure 3: Comparison of the true distribution (blue) and model's internal distributions (orange). For an ideal model, its internal distribution coincides with the true distribution.

4.2 Theoretical Analysis

Distributional Consistency We begin with the property of ideal models. Given any history $\mathbf{X}_{-H:t-1}$, TSLM generates an *internal probability distribution* f_θ. From the perspective of this distribution, we revisit the ideal models and conclude the distribution-perfect prediction property.

Lemma 4.1. *For $\sigma_t^2 \ge 0$, we have $f_\theta \equiv f_t$ a.e.*

Fig. 3 illustrates the insight of Lemma 4.1. Since the probability density of $X_t = T_t + d$ is $f_t(T_t + d)$, the loss expectation is therefore the cross entropy $-\int f_t(v) \log f_\theta(v) \mathrm{d}v$ between f_θ and the true series distribution f_t. As illustrated in Fig. 3, it is minimized when $f_\theta \equiv f_t$ almost everywhere by Gibbs' inequality. Specifically, if the true sequence is entirely deterministic (i.e., zero uncertainty), ideal models make value-perfect predictions, leading to Corollary 4.2.

Corollary 4.2. *If $\sigma_t^2 = 0$, we have $f_\theta(z = v) = \delta(v - T_t)$, where δ is the Dirac-δ function.*

In summary, the ideal model's internal distributions coincide with true series distributions and therefore preserve the inherent uncertainty in the series.

Sampling-Induced Variance Scaling Using the property of ideal models to reproduce true series distributions, we analyze the prediction of $\hat{\mathbf{X}}_{1:\tau}$ with history $\mathbf{X}_{-H:0}$, and focus on the variation in uncertainty. In this section, we explain that the sampling probability distribution $\hat{f}_t$ of each X_t is a modified version of f_θ, which tends to reduce its uncertainty.

TSLMs modify internal distributions f_θ through sampling strategies to modulate output diversity (Radford et al. 2019). We summarize common sampling strategies as follows: (1) scaling (e.g., temperature sampling); and (2) symmetric truncation (with normalization, e.g., top-k). These modifications preserve the means of f_θ but alter their variances, as shown in Fig. 4, generating sampling probability distributions $\hat{f}_\theta(\hat{X}_t = v)$ to sample $\hat{X}_t$ with variance $\hat{\sigma}_t^2$.

Let the uncertainty (variance) of f_θ be $\tilde{\sigma}_t^2$ at time t, we denote the modified uncertainty by sampling strategies as $\hat{\sigma}_t^2 = \gamma_t \cdot \tilde{\sigma}_t^2$. Intuitively, γ_t controls uncertainty expansion

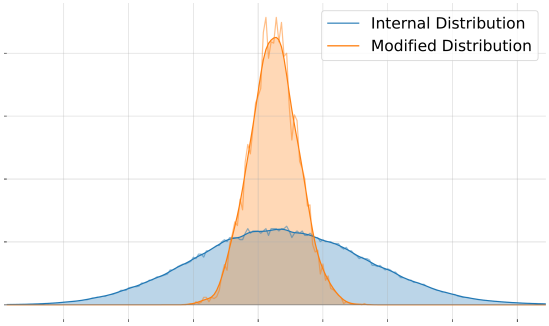


Figure 4: Comparison of the internal distribution (blue, from the model logits) and modified sampling distribution (orange, from 10,000-step Monte Carlo) for a single prediction step. The model's internal probability distribution becomes sharper under our sampling strategy.

($\gamma_t > 1$), preservation ($\gamma_t = 1$), or contraction ($\gamma_t < 1$). Since modifications of sampling strategies preserve the expectation, we recursively obtain $\mathbb{E}(\hat{X}_t) = T_t$ for all $1 \leq t \leq \tau$. We therefore treat the generation of $\hat{\mathbf{X}}_{1:\tau}$ as τ independent, one-step predictions given their respective histories. We evaluate the differences between $\mathbf{X}_{1:\tau}$ and $\hat{\mathbf{X}}_{1:\tau}$ using the model evaluation function, leading to Lemma 4.3.

Lemma 4.3. *The expectation of the model evaluation function* $\mathbb{E}(\text{Eva})$ *increases monotonically with respect to all* $\hat{\sigma}_t$.

Lemma 4.3 indicates that a smaller $\hat{\sigma}_t^2$ yields a lower Eva value, suggesting a higher quality of generation. Given $\hat{\sigma}_t^2 = \gamma_t \cdot \tilde{\sigma}_t^2$, when $\tilde{\sigma}_t^2$ is fixed, a smaller γ_t yields lower $\mathbb{E}(\text{Eva})$, which validates our intuition.

In particular, we derive a corollary from Lemma 4.3 that if metrics such as MSE are used as training loss, the contraction of uncertainty is immediate.

Corollary 4.4. *If the loss function is defined in the same form as the model evaluation function* Eva *in Eq.* (1) *(e.g., MSE),*

we have $f_\theta(z = v) = \delta(v - T_t)$.

Based on the analysis, we intuitively expect that TSLMs employ $\gamma_t < 1$ in practice. In the following, we further analyze this intuition of γ_t in terms of performance.

Recursive Variance Reduction With the intuition of γ_t above, we examine how the sampling distribution (i.e., the internal distribution modified by γ_t) affects the subsequent generation. In this section, we demonstrate the recursive evolution of the variance and analyze its behavior under different values of the scaling factor γ_t.

According to Lemma 4.1, the ideal model reproduces any distribution with variance satisfying $\sigma_t^2 = \sum_{i=1}^{l} \alpha_i \sigma_{t-i}^2$. When forecasts $\hat{X}_1, \hat{X}_2, \ldots$ serve as histories, their variances $\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots$ also follow the same recurrence relation.

$$\tilde{\sigma}_t^2 = \sum_{i=1}^{l} \alpha_i \hat{\sigma}_{t-i}^2 = \sum_{i=1}^{l} \alpha_i \gamma_{t-i} \tilde{\sigma}_{t-i}^2, \sum_{i=1}^{l} \alpha_i = 1. \quad (2)$$

Typically, the uncertainty scaling behavior is consistent and $\gamma_t < 1$ or $\gamma_t > 1$ for all t. When $\gamma_t < 1$, the internal uncertainty, according to Eq. (2), is strictly lower than the weighted sum of historical uncertainties, leading to an exponential decay in uncertainty towards 0. By Corollary 4.2, the forecast series $\hat{X}_t$ converges to the trend T_t. Conversely, when $\gamma_t > 1$, the uncertainty grows exponentially towards infinity, introducing excessive noise into the forecast that undermines trend capture and violates forecasting objectives. When $\gamma_t \equiv 1$, the uncertainty in forecast series is relatively stable for ideal models, which is unattainable for practical models due to noise accumulation.

This theoretical analysis also helps explain why many practical TSLMs exhibit similar behaviors of uncertainty contraction. For example, Chronos (Ansari et al. 2024) adopts contractive sampling strategies by using top-k with median sampling, and therefore $\gamma_t < 1$ holds consistently. Timer (Liu et al. 2024) and Time-MoE (Shi et al. 2025) utilize point-wise errors (e.g., MSE) as training loss, sharing the form as Eva defined in Eq. (1). According to Corollary 4.4, the forecast uncertainty converges rapidly, indicating their equivalence to $\gamma_t < 1$ in terms of performance. We formalize the aforementioned analysis as *contraction hypothesis*, where variance can be substituted for other metrics.

Theorem 4.5. *For an Ideal Model's forecast time series* $\hat{\mathbf{X}}_t$ *with history* $\mathbf{X}_{-H:0}$*, let the internal probability distribution at* t *be* $f_\theta(z_t)$*, and define its uncertainty measure by variance* $\tilde{\sigma}_t^2$*. Therefore we have that* $\tilde{\sigma}_t^2$ *is monotonically decreasing with* t *and* $\lim_{t\to\infty} \tilde{\sigma}_t^2 \approx 0$.

We provide empirical verification of Theorem 4.5 in Figure 2, which illustrates this uncertainty contrast over 1024 tokens using three metrics, namely entropy, max-probability, and variance, which are used in our detection method. In model-generated series, max-probability steadily approaches 1, while both entropy and variance decay towards 0. In contrast, real sequences remain comparatively stable in uncertainty. This observed discrepancy in series uncertainty, combined with the ability of TSLMs to capture it, forms the basis of our detection methodology.

4.3 Uncertainty Contraction Estimator

The Uncertainty Contraction Estimator (UCE) operationalizes the Contraction Hypothesis by quantifying the uncertainty contraction in model-generated time series. Given a candidate $\mathbf{X}_t$, we first sample N time points $t_1, t_2, \ldots, t_N$ with a fixed interval $\Delta t = t_{i+1} - t_i$. For each t_i, UCE takes $\mathbf{X}t_i$ as the input to the TSLM and computes its internal probability distribution $\hat{P}_{t_i}$ as defined in Eq. (3), yielding a sequence of internal distributions.

$$\hat{P}_{t_i} = p_\theta\left(\cdot | X_1, \ldots, X_{t_i}\right), i = 1, \ldots, N. \quad (3)$$

By Lemma 4.1 the internal probability distribution coincides with the underlying Gaussian noise distribution, which is unimodal and relatively probability-concentrated. Therefore, UCE focuses on a neighborhood $\mathcal{U}$ around the mean. Within $\mathcal{U}$, UCE computes three uncertainty measures to capture different aspects of the distribution concentration.

1. *Entropy* $E = -\sum_{x\in\mathcal{U}} \hat{P}(x) \log \hat{P}(x)$, which captures the spread of probability mass.

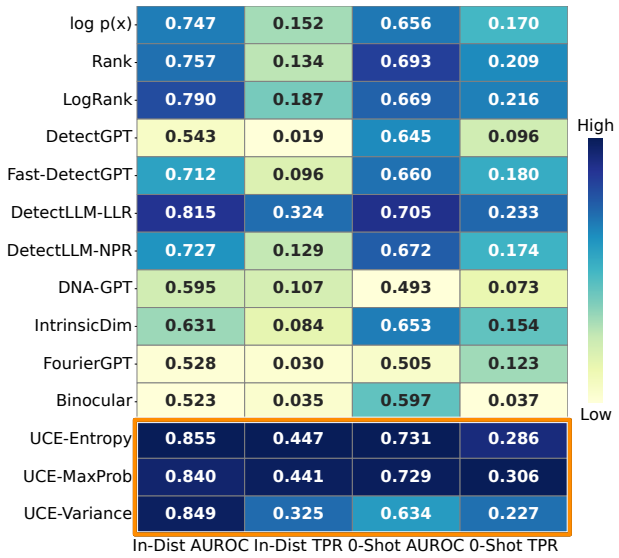


Figure 5: Average AUROC and TPR (at 1% FPR) for model generation detection on In-Distribution (12 datasets) and Zero-Shot (20 datasets) scenarios.

2. *Max-Probability* $P_{\max} = \max_{x \in \mathcal{U}} \hat{P}(x)$, where larger values imply lower uncertainty.
3. *Variance* $\mathrm{Var} = \sum_{x \in \mathcal{U}} (x - \mu)^2 \hat{P}(x)$ with μ the local mean to quantify concentration.

For a selected metric $s \in \{E, P_{\max}, \mathrm{Var}\}$, UCE calculates the metric sequence $s_{t_1}, s_{t_2}, \ldots, s_{t_N}$, and the overall UCE score is their mean, formulated as $\mathrm{UCE} = \frac{1}{N} \sum_{i=1}^{N} s_{t_i}$.

5 Experiments

5.1 Experimental Settings

Model and Datasets We evaluate all detection methods using Chronos-T5 (large), a point-forecasting TSLM that discretizes continuous values over a fixed token vocabulary and produces probability distributions. We conduct experiments on 32 datasets across diverse domains (energy, finance, transportation), categorized as (1) in-distribution datasets, including 12 used during Chronos training; and (2) zero-shot datasets, consisting of 20 previously unseen datasets. The zero-shot setting evaluates the generalization of detection to unseen datasets to eliminate the effects of data leakage, and validates the "black-box" detection capability (see Section 5.3). For each dataset, we follow the default setup in Chronos to generate forecasts of horizon $H = 64$, and use the corresponding H observations as ground truth. We treat forecast series as positive (model-generated) samples and real series as negative samples.

Baselines We compare UCE with 10 text-based baselines adapted for time series: (1) **DNA-GPT WScore** (Yang et al. 2024), which measures the average log-likelihood gap between regenerated and original text given a fixed prefix; (2) **DetectGPT** (Mitchell et al. 2023) and its efficient variant **Fast-DetectGPT** (Bao et al. 2023), which detect generation by measuring probability changes under model-guided perturbations; (3) **DetectLLM** (Su et al. 2023), using Log-Likelihood Log-Rank Ratio (LRR) and Normalized Perturbed log-Rank (NPR) metrics; (4) **Intrinsic Dimension** (Tulchinskii et al. 2023), which classifies sequences by estimating the topological "intrinsic dimension" of token embeddings; (5) **FourierGPT** (Xu et al. 2024), performing spectral analysis on token probability sequences; (6) **Binocular** (Hans et al. 2024), leveraging two models to reduce the intrinsic perplexity of prefixes; and (7) **Traditional metrics** (Gehrmann, Strobelt, and Rush 2019; Jawahar, Abdul-Mageed, and Lakshmanan 2020; Solaiman et al. 2019), including log-likelihood $\log p(x)$, rank, and log-rank. The details of the baselines are documented in the Appendix.

Evaluation Metrics We use the Area Under the ROC Curve (AUROC) to evaluate the detection performance of UCE and the baseline methods. Since AUROC masks performance at low false positive rates (FPR), which are critical in generation detection tasks (Carlini et al. 2022; Krishna et al. 2023; Yang et al. 2024), we also report the true positive rate (TPR) at a fixed FPR of 1% (Roberts et al. 2024).

5.2 Overall Result

We evaluate detection performance under in-distribution and zero-shot settings. UCE consistently achieves state-of-the-art AUROC and TPR across all scenarios, with key results shown in Fig. 5.

In-Distribution UCE-Entropy achieves an AUROC of 0.855, outperforming the strongest baseline, DetectLLM-LLR (0.815), by 0.040, and exceeding the baseline average (0.670) by 0.183. Its TPR reaches 0.447, surpassing DetectLLM-LLR (0.324) by 0.123 and the baseline average (0.118) by 0.329.

Zero-Shot In the zero-shot setting, UCE-Entropy achieves an AUROC of 0.731, outperforming DetectLLM-LLR (0.705) by 0.026 and the baseline average (0.632) by 0.099. It also attains a TPR of 0.286, exceeding DetectLLM-LLR (0.233) by 0.053 and the average baseline (0.151) by 0.135.

The other two UCE metrics (Max-Probability and Variance) also show strong performance, with UCE–MaxProb approaching UCE–Entropy in TPR despite a modest AUROC gap; UCE–Variance is slightly weaker in TPR. Overall, UCE-Entropy remains the top performer, UCE-MaxProb is a close second, and UCE-Variance may offer greater robustness under alternative conditions. The full experimental results, with additional experiments, are provided in the Appendix. The experimental results suggest potential future work on hybrid or adaptive variant selection.

5.3 Cross-Model Detection Performance of UCE

In this section, we evaluate the generalization of UCE on time series generated by two alternative TSLMs, Timer and Time-MoE. We conduct experiments on 9 datasets covering multiple lengths and report both AUROC and TPR.

UCE achieves strong overall performance on both models, with the Entropy variant exhibiting particularly notable results, consistently achieving high AUROC and TPR across

Time Series Length H	UCE-Entropy		UCE-Max Prob		UCE-Variance	
	AUROC	TPR	AUROC	TPR	AUROC	TPR
$H = 96$	0.833	0.301	0.556	0.041	0.635	0.108
$H = 192$	0.771	0.280	0.508	0.050	0.576	0.102
$H = 336$	0.765	0.305	0.484	0.032	0.564	0.093
$H = 768$	0.788	0.366	0.542	0.096	0.602	0.179

Table 2: AUROC and TPR (at 1% FPR) for Timer generation detection on 9 datasets.

Time Series Length H	UCE-Entropy		UCE-Max Prob		UCE-Variance	
	AUROC	TPR	AUROC	TPR	AUROC	TPR
$H = 96$	0.829	0.320	0.717	0.292	0.745	0.316
$H = 192$	0.890	0.392	0.773	0.420	0.806	0.422
$H = 336$	0.957	0.611	0.810	0.475	0.856	0.511
$H = 720$	0.950	0.561	0.845	0.540	0.863	0.566

Table 3: AUROC and TPR (at 1% FPR) for Time-MoE generation detection on 9 datasets.

both models, as shown in Table 2 and Table 3. The Max-Probability and Variance variants yield weaker performance, especially for Timer-generated sequences. Notably, UCE performs particularly well in Time-MoE, possibly due to its Mixture-of-Experts architecture with better long-forecasting performance. In Section 4.2 we show that although Timer and Time-MoE are not probabilistic forecasting models, they also exhibit uncertainty contraction, and uncertainty metrics demonstrate discriminative power between real and model-generated series.

Furthermore, the performance of UCE on both zero-shot datasets and cross-model settings also implies its potential for the "black-box" detection analogous to DetectGPT. Specifically, given a time series that originates from an unknown model and unknown source (or the real world), UCE can perform detection by leveraging a locally deployed probabilistic model to compute uncertainty-based signals.

6 Discussion

Our methodology is based on the contraction hypothesis, which depends on certain idealized assumptions (see Section 4.1). Despite the empirical support for the hypothesis from the experimental results, we revisit the foundational assumptions to strengthen our theoretical analysis.

6.1 Idealized Model Assumption

To establish an architecture-agnostic detection methodology, we postulate a theoretically optimal Ideal Model. This abstraction avoids model-specific details to identify general behavior, but is mathematically unrealizable in practice. Existing TSLMs are approximations of this ideal, which undermine both the guarantee of perfect prediction (see Corollary 4.2) and the model's ability to faithfully recover the true series distribution (see Lemma 4.1). Despite the limitation, practical models also exhibit similar uncertainty contraction behavior in forecasting.

In particular, TSLMs are trained on finite datasets with finite parameters. Since the noise samples are limited and the model is trained to minimize expected loss over these few realizations, it cannot fully capture the underlying noise distribution or the true trend. Therefore, both the estimated trend $\hat{T}_t$ and noise $\hat{n}_t$ inevitably deviate from the truth, and the trend deviation $\Delta T_t = |T_t - \hat{T}_t|$ accumulates as forecasting progresses. Given this non-zero deviation, Lemma 4.3 does not strictly hold, and reducing forecast variance $\hat{\sigma}_t^2$ does not always improve evaluation performance. Nevertheless, a moderate reduction in uncertainty constrains the expected evaluation error within the relatively small bound ΔT_t by concentrating probability between T_t and $\hat{T}_t$. As the deviation grows, greater uncertainty does not yield sufficiently better performance by correcting the error and may result in unstable forecasts. This indicates that practical TSLMs also inherently exhibit a forecasting uncertainty contraction (i.e., $\gamma_t < 1$), especially under recursive noise accumulation, as elaborated in Section 4.2. Hence, the contraction hypothesis still holds for practical models, allowing UCE to track predictive uncertainty across recursive steps.

6.2 Gaussian Noise Assumption

In this work, we assume Gaussian noise for analytical simplification: the sum of independent Gaussian random variables remains Gaussian, and the variance fully characterizes the distribution. Crucially, our proofs do not depend on the exact form of the Gaussian but only require the noise to be unimodal (greatest probability for trend), symmetric (equivalence for both positive and negative biases), and having a finite second moment (existence of variance). By the additivity of the variance for independent variables, our results naturally extend to such noise models. These properties are shared by most noise models commonly assumed in real-world time series analysis (e.g., Laplace), and the experimental results (see Appendix) validate the broad applicability of the contraction hypothesis under various noise types.

6.3 Modality Difference Analysis

In this work, we extend textual detection methods to time series via UCE. Although uncertainty metrics (e.g., entropy) are relatively simple, we investigate the unique properties of time series versus text modalities (see Section 4.2) to prove the optimality of uncertainty in discriminative power.

7 Conclusion

We investigate detecting TSLM-generated time series and hypothesize that they exhibit progressively decreasing uncertainty—unlike real data. Building on this, we propose the Uncertainty Contraction Estimator (UCE), which captures uncertainty to distinguish model-generated from real series and is validated both theoretically and empirically. Future work will further analyze the hypothesis under diverse model architectures and extend UCE to multivariate and batch-forecasting settings.

Acknowledgements

This work was supported in part by National Key Research and Development Plan in China (2023YFC3306100) and National Natural Science Foundation of China (62272372).

References

Ansari, A. F.; Stella, L.; Turkmen, C.; Zhang, X.; Mercado, P.; Shen, H.; Shchur, O.; Rangapuram, S. S.; Pineda Arango, S.; Kapoor, S.; Zschiegner, J.; Maddix, D. C.; Mahoney, M. W.; Torkkola, K.; Gordon Wilson, A.; Bohlke-Schneider, M.; and Wang, Y. 2024. Chronos: Learning the Language of Time Series. *Transactions on Machine Learning Research*.

Bao, G.; Zhao, Y.; Teng, Z.; Yang, L.; and Zhang, Y. 2023. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. In *Proceedings of the International Conference on Learning Representations*.

Box, G. E. P.; Jenkins, G. M.; Reinsel, G. C.; and Ljung, G. M. 1978. *Time Series Analysis: Forecasting and Control*. Wiley.

Brumfiel, G. 2002. Bell labs launches inquiry into allegations of data duplication. *Nature*, 417(6892): 367–368.

Carlini, N.; Chien, S.; Nasr, M.; Song, S.; Terzis, A.; and Tramèr, F. 2022. Membership Inference Attacks From First Principles. In *IEEE Symposium on Security and Privacy*, 1897–1914.

Gehrmann, S.; Strobelt, H.; and Rush, A. M. 2019. GLTR: Statistical Detection and Visualization of Generated Text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 111–116.

Godlee, F.; Smith, J.; and Marcovitch, H. 2011. Wakefield's article linking MMR vaccine and autism was fraudulent. *BMJ (Clinical research ed.)*, 342: c7452.

Goswami, M.; Szafer, K.; Choudhry, A.; Cai, Y.; Li, S.; and Dubrawski, A. 2024. MOMENT: A Family of Open Time-series Foundation Models. In *Proceedings of the International Conference on Learning Representations*.

Gupta, M. 2024. Defending Against LLM-Based Financial Fraud: Best Practices and Recommendations. Accessed: 2025-07-08.

Hans, A.; Schwarzschild, A.; Cherepanova, V.; Kazemi, H.; Saha, A.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2024. Spotting LLMs with binoculars: zero-shot detection of machine-generated text. In *Proceedings of the International Conference on Machine Learning*.

Hashimoto, T. B.; Zhang, H.; and Liang, P. 2019. Unifying Human and Statistical Evaluation for Natural Language Generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1689–1701.

Jawahar, G.; Abdul-Mageed, M.; and Lakshmanan, L. 2020. Automatic Detection of Machine Generated Text: A Critical Survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, 2296–2309.

Jin, M.; Wang, S.; Ma, L.; Chu, Z.; Zhang, J. Y.; Shi, X.; Chen, P.; Liang, Y.; Li, Y.; Pan, S.; and Wen, Q. 2024. Time-LLM: Time series forecasting by reprogramming large language models. In *Proceedings of the International Conference on Learning Representations*.

Krishna, K.; Song, Y.; Karpinska, M.; Wieting, J.; and Iyyer, M. 2023. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems*, volume 36, 27469–27500.

Lavergne, T.; Urvoy, T.; and Yvon, F. 2008. Detecting fake content with relative entropy scoring. In *Proceedings of the 2008 International Conference on Uncovering Plagiarism, Authorship and Social Software Misuse*, volume 377, 27––31.

Liu, Y.; Zhang, H.; Li, C.; Huang, X.; Wang, J.; and Long, M. 2024. Timer: Generative Pre-trained Transformers Are Large Time Series Models. In *Proceedings of the International Conference on Machine Learning*.

Meister, C.; Wiher, G.; Pimentel, T.; and Cotterell, R. 2022. On the probability–quality paradox in language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 36–45.

Mitchell, E.; Lee, Y.; Khazatsky, A.; Manning, C. D.; and Finn, C. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. In *Proceedings of the International Conference on Machine Learning*, volume 202, 24950–24962.

Nate, G.; Marc, F.; Qiu, S.; and Wilson, A. G. 2023. Large Language Models Are Zero-Shot Time Series Forecasters. In *Advances in Neural Information Processing Systems*.

Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2023. A Time Series Is Worth 64 Words: Long-Term Forecasting with Transformers. In *Proceedings of the International Conference on Learning Representations*.

Pu, J.; Sarwar, Z.; Abdullah, S. M.; Rehman, A.; Kim, Y.; Bhattacharya, P.; Javed, M.; and Viswanath, B. 2023. Deepfake Text Detection: Limitations and Opportunities. In *Proc. of IEEE S&P*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Roberts, M.; Hazan, A.; Dittmer, S.; Rudd, J. H. F.; and Schönlieb, C.-B. 2024. The curious case of the test set AUROC. *Nature Machine Intelligence*, 6(4): 373–376.

Rose, C. S.; and Sesia, A. 2013. Barclays and the LIBOR Scandal. Technical Report 313-075, Harvard Business School.

Shi, X.; Wang, S.; Nie, Y.; Li, D.; Ye, Z.; Wen, Q.; and Jin, M. 2025. Time-MoE: Billion-Scale Time Series Foundation Models with Mixture of Experts. In *International Conference on Representation Learning*, volume 2025, 34635–34667.

Solaiman, I.; Brundage, M.; Clark, J.; Askell, A.; Herbert-Voss, A.; Wu, J.; Radford, A.; and Wang, J. 2019. Release Strategies and the Social Impacts of Language Models. *ArXiv*.

Su, J.; Zhuo, T. Y.; Wang, D.; and Nakov, P. 2023. DetectLLM: Leveraging Log Rank Information for Zero-Shot Detection of Machine-Generated Text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 12395–12412.

Tan, M.; Merrill, M. A.; Gupta, V.; Althoff, T.; and Hartvigsen, T. 2024. Are Language Models Actually Useful for Time Series Forecasting? In *Advances in Neural Information Processing Systems*.

Tulchinskii, E.; Kuznetsov, K.; Kushnareva, L.; Cherniavskii, D.; Nikolenko, S.; Burnaev, E.; Barannikov, S.; and Piontkovskaya, I. 2023. Intrinsic Dimension Estimation for Robust Detection of AI-Generated Texts. In *Advances in Neural Information Processing Systems*, 39257–39276.

Uchendu, A.; Le, T.; Shu, K.; and Lee, D. 2020. Authorship Attribution for Neural Text Generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 8384–8395.

Wang, X.; Wang, X.; and Wilkes, M. 2021. *Unsupervised Fraud Detection in Environmental Time Series Data*, 257–277. Springer Singapore.

Xu, Y.; Wang, Y.; An, H.; Liu, Z.; and Li, Y. 2024. Detecting Subtle Differences between Human and Model Languages Using Spectrum of Relative Likelihood. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 10108–10121.

Xue, H.; and Salim, F. D. 2023. PromptCast: A New Prompt-Based Learning Paradigm for Time Series Forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 1–14.

Yang, X.; Cheng, W.; Wu, Y.; Petzold, L. R.; Wang, W. Y.; and Chen, H. 2024. DNA-GPT: Divergent N-Gram Analysis for Training-Free Detection of GPT-Generated Text. In *Proceedings of the International Conference on Learning Representations*.

Yao, Q.; Yang, C.-H. H.; Jiang, R.; Liang, Y.; Jin, M.; and Pan, S. 2025. Towards Neural Scaling Laws for Time Series Foundation Models. In *Proceedings of the International Conference on Learning Representations*.

Zhou, T.; Niu, P.; Wang, X.; Sun, L.; and Jin, R. 2023. One Fits All: Power General Time Series Analysis by Pretrained LM. In *Advances in Neural Information Processing Systems*, 43322–43355.