Speaking Numbers to LLMs: Multi-Wavelet Number Embeddings for Time Series Forecasting

Defu Cao[1] , **Zijie Lei**[1] , **Muyan Weng**[1] , **Jiao Sun**[1,2] and **Yan Liu**[1]
[1]University of Southern California
[2]Google DeepMind
{defucao, zijielei, muyanwen, jiaosun, yanliu.cs}@usc.edu

Abstract

Large language models (LLMs) are attractive for context-aware time series forecasting because they can integrate heterogeneous textual signals, yet their discrete, language-oriented tokenization and embedding interfaces are misaligned with continuous numerical values, often harming numerical ordering and forecasting reliability. We propose TempoWave, a plug-and-play temporal wavelet digit interface that maps each scalar observation into digit-wise embeddings constructed from multi-wavelet, multi-scale coefficients. By directly overriding standard token representations, TempoWave seamlessly exposes both fine-grained local fluctuations and macro global structures in a transformer-compatible form, ensuring that precise numerical formatting, distinct digit identity, and robustness to common normalization operations are maintained throughout the LLM pipeline. Experiments across five context-enriched forecasting benchmarks demonstrate that TempoWave consistently improves LLM-based forecasters over standard numeric tokenization and alternative embedding interfaces, achieving a new state-of-the-art. These results highlight the numeric interface as a key bottleneck and suggest that principled multi-resolution embeddings can better couple LLMs' contextual reasoning with precise forecasting. Our code is available at DC-research/TempoWAVE and our model can be accessed at Melady/TempoWAVE.

1 Introduction

Time series analysis, the study of data points ordered chronologically, is indispensable across sectors such as finance, healthcare, and climate science [Cao *et al.*, 2025; Wang *et al.*, 2026b]. Accurate forecasting supports resource allocation, risk management, and early warning systems, yet the underlying data generating processes are often complex and evolving [Cao *et al.*, 2023a; Zhang *et al.*, 2022]. Practical time series typically exhibit non-stationarity, mixed periodicities, regime shifts, long- and short-range temporal dependencies, and substantial noise [Liu *et al.*, 2022; Cao *et al.*, 2020; Cao *et al.*, 2021; Cao *et al.*, 2023b]. These properties make it difficult to learn models that simultaneously capture fine-grained local fluctuations and long-horizon global structure, while remaining robust under distribution shifts and limited supervision.

In parallel, Large Language Models (LLMs) [OpenAI, 2023] have become strong general-purpose sequence learners. They can exploit long contexts, perform in-context pattern induction, and naturally integrate textual information [Hu *et al.*, 2025; Zhou and Yu, 2025; Zhang *et al.*, 2024]. This is particularly appealing for time series intelligence because many exogenous drivers that affect temporal dynamics are expressed in language, such as policy changes, market news, clinical narratives, and operational logs. Moreover, the few-shot and zero-shot generalization behavior of LLMs suggests a promising pathway for domains where labeled time series data is scarce and task distribution varies across entities, locations, or time periods.

Despite this promise, directly adapting LLMs to time series forecasting remains challenging. LLMs are optimized for discrete token prediction, whereas time series forecasting fundamentally requires precise modeling of continuous values. This mismatch can lead to unreliable numerical behavior even when the model captures high-level temporal patterns [Merrill *et al.*, 2024; Ye *et al.*, 2025]. More critically, language-oriented tokenization fragments numbers into sub-tokens in ways that are not tied to magnitude, for example "2026" $\rightarrow$ "20" and "26". Such fragmentation breaks ordinal relations and obscures the continuity intrinsic to temporal processes. As a result, two numerically close values may be mapped to very different token sequences, while numerically distant values can share sub-tokens, introducing spurious similarity. In LLM-based forecasting pipelines, this translation layer between real-valued sequences and discrete tokens becomes a principal bottleneck.

Recent research has pursued several avenues to bridge the gap between LLMs and time series analysis. One direction develops specialized foundation models tailored to time series [Cao *et al.*, 2026; Cao *et al.*, 2024a]. Another uses agentic or multimodal systems that couple LLMs with dedicated forecasting tools [Ye *et al.*, 2026; Li *et al.*, 2026]. A third direction focuses on input adaptations, including patching [Nie *et al.*, 2023], quantization [Talukder *et al.*, 2024], or converting time series into symbolic or textual representations [Jia *et*

al., 2024]. While these approaches can improve usability and efficiency, they often trade away numerical faithfulness, blur fine-grained variations, or rely on external components that reduce end-to-end differentiability and complicate analysis. Consequently, there remains a persistent gap in representing continuous numerical values inside the standard transformer input space in a principled and information-preserving manner.

In this work, we focus on the representation bottleneck and ask whether LLMs can be equipped with numerically grounded embeddings that preserve quantitative relations and multi-scale temporal structure while remaining compatible with standard transformer inputs. An effective forecasting representation should support reasoning across resolutions. Local changes are crucial for short-term dynamics and anomaly-sensitive regimes, while trends and seasonal components dominate long-horizon behavior. Motivated by wavelet analysis, which provides a natural multi-resolution decomposition, we propose *Multi-Wavelet Number Embedding (TempoWave)*, an input embedding interface that maps each scalar observation into a dense vector encoding multi-scale structure. TempoWave is designed to be injected into LLM backbones without requiring language tokenization of numbers, thereby reducing the disconnect between numeric magnitude and the model's discrete interface. Beyond forecasting accuracy, we also aim to understand how the embedding interface shapes numerical structure inside the model. To this end, we introduce diagnostic analyses that probe whether local neighborhoods in representation space respect numeric ordering, a property we refer to as monotonic neighborhood consistency. These analyses help explain when TempoWave improves forecasting and provide guidance for designing numerically grounded interfaces for LLMs. Extensive experiments on diverse forecasting benchmarks show that TempoWave consistently improves LLM-based forecasters compared to standard tokenization and common adaptation methods, and it is competitive with strong time-series-specific models in settings requiring precise numerical forecasting.

Contributions. Our contributions are as follows:

- **A wavelet-based numeric interface for LLM forecasting.** We propose *Multi-Wavelet Number Embedding (TempoWave)*, an input embedding interface that maps each real-valued numerical observation into a dense vector with multi-resolution structure, enabling direct use of standard LLM backbones for numerical forecasting without relying on language tokenization of numbers.
- **Structural Faithful and Stable Numeric Representation.** We analyze TempoWave from a multi-scale signal processing perspective and establish properties related to numerical faithfulness and stability, including improved separability across values and robustness to common normalization operations used in transformers.
- **Comprehensive evaluation with diagnostic evidence.** We conduct extensive experiments on diverse forecasting benchmarks and show consistent gains over LLM baselines using tokenization and common input adaptations. We further provide diagnostic analyses that probe monotonic neighborhood consistency, offering evidence for how TempoWave reshapes numerical neighborhoods and helping explain its forecasting improvements.

2 Related Work

Research on applying Large Language Models (LLMs) to time series analysis has grown rapidly, motivated by LLMs' strong sequence modeling and their ability to integrate contextual information. Existing efforts can be grouped into three directions: time series foundation models trained on large-scale temporal corpora, LLM-centered agentic or multimodal systems, and input adaptation strategies that re-design how numerical values are represented for transformer backbones.

Time series foundation models. A major direction is to pre-train foundation models directly on large and diverse time series collections to learn transferable temporal representations. Representative examples include Chronos [Ansari *et al.*, 2024], TimesFM [Das *et al.*, 2024], and other large-scale temporal pre-training efforts [Woo *et al.*, 2024; Cao *et al.*, 2024b; Yang *et al.*, 2025a]. These models demonstrate the benefits of scaling and pre-training for forecasting, but they often rely on patching, discretization, or quantization to interface with transformers, which can blur fine-grained numerical differences and introduce information loss in precision-sensitive regimes.

LLM agents and multimodal time series systems. Another line of work uses general-purpose LLMs as reasoning engines within larger pipelines, where forecasting or numerical computation is delegated to specialized tools and the LLM performs orchestration and interpretation [Yang *et al.*, 2026; Weng *et al.*, 2026; Yang *et al.*, 2025b]. Closely related are multimodal systems that jointly model time series and text, enabling context-aware forecasting and question answering, for example TimeLLM [Jin *et al.*, 2024], GPT4MTS [Jia *et al.*, 2024]. While these approaches highlight the value of fusing textual signals, their performance still depends critically on how continuous values are encoded for transformer inputs, and numerical faithfulness can remain a bottleneck when the interface is token-based or heavily discretized.

Input adaptation and numeric representation for LLMs. A substantial body of work focuses on making numerical time series consumable by LLM backbones via input adaptation. Common strategies include patch-based representations [Nie *et al.*, 2023], discretization or binning [Talukder *et al.*, 2024], symbolic conversion of segments into strings [Goswami *et al.*, 2024], and embedding alignment methods that map time series embeddings into the language embedding space [Gruver *et al.*, 2024; Zeng *et al.*, 2023]. More generally, recent studies on numeracy in language models emphasize that tokenization and discrete interfaces can impede faithful numerical reasoning, motivating alternative representations that preserve quantitative structure [Merrill *et al.*, 2024; Gillman *et al.*, 2025]. These adaptations improve compatibility, but they often shift the burden to a preprocessing stage and may sacrifice either precision, locality, or multi-scale structure [Wang *et al.*, 2026a].

Positioning of TempoWave. Our work addresses the above interface challenge by introducing Multi-Wavelet Number

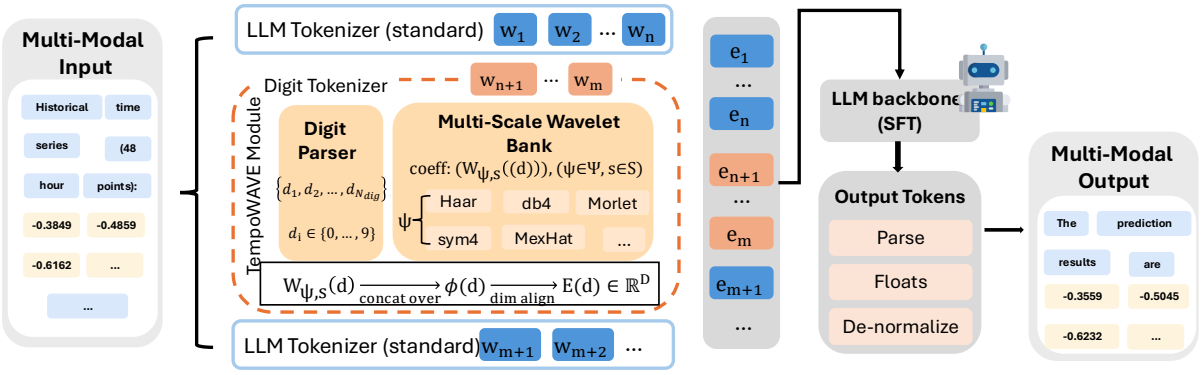

Figure 1: Overview of the TempoWave-based forecasting framework with digit-level tokens. The input prompt is tokenized once using a tokenizer augmented with dedicated digit tokens. Text and context tokens use standard embeddings, while digit tokens are routed to the TempoWave module, which constructs digit embeddings via multi-wavelet, multi-scale coefficients and overrides the corresponding token embeddings. The resulting embedding sequence is fed into an unchanged LLM backbone trained via supervised fine-tuning (SFT). The model generates numeric tokens that are parsed, de-normalized, and evaluated as real-valued forecasts.

Embedding, which constructs numerically grounded embeddings that encode multi-resolution structure prior to ingestion by an LLM. In contrast to purely symbolic conversion or coarse discretization, TempoWave aims to preserve quantitative relations while providing a multi-scale representation inspired by wavelet analysis, supporting both accurate forecasting and diagnostic analysis of the induced numerical neighborhood structure.

3 Methodology

3.1 Overview

To bridge the mismatch between continuous-valued time series and the discrete input interface of Large Language Models (LLMs), we propose Multi-Wavelet Number Embedding (TempoWave), a numerically grounded embedding interface that intervenes only at the token embedding layer while keeping the LLM backbone unchanged. As illustrated in Figure 1, TempoWave enables LLMs to process numerical sequences by replacing the embeddings of digit tokens with multi-resolution wavelet-based representations.

Numeric-to-token interface. Given a normalized time series value $x_t \in \mathbb{R}$, we first render it into a fixed-precision string with m_{prec} integer digits and n_{prec} fractional digits (e.g., `V.FFFF`). A single tokenization pass is then applied using a tokenizer augmented with *dedicated digit tokens*, where each digit $d_i \in \{0, \ldots, 9\}$ is treated as an individual token. As a result, the input prompt is converted into a mixed token sequence consisting of text/context tokens and digit tokens.

Standard token embeddings are used for text and context tokens. For digit tokens, TempoWave computes digit embeddings via multi-wavelet, multi-scale features and *overrides* the standard embeddings. This routing mechanism ensures that numerical structure is injected at digit positions only, without altering the remaining token representations.

TempoWave embedding override. For each digit token d, TempoWave computes a set of wavelet coefficients $W_{\psi,s}(d)$ over a predefined wavelet family Ψ and scale set S. These coefficients are concatenated into a digit feature vector $\phi(d)$ and mapped to the LLM embedding dimension via a fixed alignment function $g(\cdot)$, producing the digit embedding $E(d) \in \mathbb{R}^D$. The resulting digit embeddings replace the standard embeddings at digit positions, yielding the final input embedding sequence $\mathbf{H}_0 \in \mathbb{R}^{T \times D}$, which is fed into the LLM.

Context and training objective. For context-aware forecasting, we fine-tune the LLM via supervised fine-tuning (SFT) on prompts that include (i) historical numeric values represented as fixed-precision digit tokens, (ii) optional global descriptors such as Catch22 features [Lubba *et al.*, 2019], and (iii) situational context such as date and domain information. The LLM backbone remains unchanged, and training minimizes the standard next-token cross-entropy loss to generate future numeric tokens corresponding to the next k time steps.

3.2 TempoWave Construction

Wavelet dictionary. Let $\psi(t)$ denote a mother wavelet and $\psi_{s,\tau}(t)$ denote its scaled and translated version:

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-\tau}{s}\right), \tag{1}$$

where $s > 0$ is the scale and τ is the translation. We select a set of wavelets $\Psi = \{\psi_1, \ldots, \psi_k\}$ and a set of scales $S = \{s_1, \ldots, s_l\}$. For TempoWave we use a fixed translation (typically $\tau = 0$).

Digit signal and wavelet coefficients. Each digit $d \in \{0, \ldots, 9\}$ is normalized to $\tilde{d} = d/9 \in [0, 1]$. To obtain wavelet coefficients without degeneracy from the zero-mean property of admissible wavelets, we represent a digit as a discrete impulse on a fixed grid. Let B be the grid resolution and $t_r = \frac{r}{B-1}$ for $r = 0, \ldots, B-1$. Define the digit index $q(d) = \lfloor \tilde{d}\,(B-1) \rceil$, and the digit signal $f_d \in \mathbb{R}^B$ as a Kronecker delta:

$$(f_d)_r = \begin{cases} 1, & r = q(d), \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

For each wavelet ψ_i at scale s_j, we sample $\psi_{i,s_j,0}(t)$ on the same grid to obtain a discrete vector $\boldsymbol{\psi}_{i,s_j} \in \mathbb{R}^B$, and define the digit wavelet coefficient as

$$W_{\psi_i,s_j}(d) := \langle f_d, \boldsymbol{\psi}_{i,s_j} \rangle = (\boldsymbol{\psi}_{i,s_j})_{q(d)}. \quad (3)$$

This definition is consistent with the continuous formulation using an impulse and avoids the trivial zero coefficients produced by projecting constants onto zero-mean wavelets.

Digit embedding and dimension matching. We concatenate multi-wavelet, multi-scale coefficients into a feature vector

$$\phi(d) = \mathrm{vec}\left(\left[W_{\psi_i,s_j}(d) \right]_{i=1..k,\, j=1..l} \right) \in \mathbb{R}^{kl}. \quad (4)$$

To interface with an LLM of embedding dimension D, we map $\phi(d)$ to a token embedding $E(d) \in \mathbb{R}^D$ via a fixed mapping $g(\cdot)$, which can be zero-padding when $kl \leq D$ or a lightweight linear projection when $kl \neq D$:

$$E(d) = g(\phi(d)) \in \mathbb{R}^D. \quad (5)$$

Since there are only ten digits, $E(0), \ldots, E(9)$ can be precomputed and cached as a small embedding table.

TempoWave for a real number and injection into LLMs. Let x be formatted into a fixed-precision digit sequence $(d_1, \ldots, d_{N_{dig}})$ with $N_{dig} = m_{prec} + n_{prec}$. TempoWave represents x as a sequence of digit token embeddings

$$\mathrm{TempoWave}(x) = \left[E(d_1), E(d_2), \ldots, E(d_{N_{dig}}) \right]. \quad (6)$$

In the input prompt, each digit token is embedded by $E(d_i)$, while other tokens use the original LLM embedding lookup. Standard positional encodings are applied as usual.

Summary on generation algorithm. Given x, we (1) extract digits according to (m_{prec}, n_{prec}), (2) compute $\phi(d)$ by evaluating wavelet samples at $q(d)$ for each (ψ_i, s_j), (3) obtain $E(d)$ via $g(\cdot)$, and (4) assemble the digit-embedding sequence $\mathrm{TempoWave}(x)$.

3.3 Representation Faithfulness in LLMs

TempoWave is designed to provide a faithful and stable numeric interface for large language models by explicitly accounting for both the discrete nature of digits and the architectural properties of Transformers. A central challenge in this setting is the pervasive use of normalization layers, such as LayerNorm and RMSNorm, which rescale and re-center token embeddings at every layer. When numerical information is encoded primarily through absolute magnitudes, such normalization can severely distort or even collapse numerical distinctions, especially under deep stacking and autoregressive decoding.

The key design principle of TempoWave is to encode digits through *structured multi-scale patterns* rather than raw scalar values. Each digit is mapped to a vector of wavelet coefficients across multiple wavelet families and scales, capturing characteristic geometric patterns in the coefficient space. By concatenating these coefficients and applying a fixed dimension-alignment mapping, TempoWave constructs digit embeddings whose identity is determined by relative patterns instead of absolute scale. As a result, subsequent normalization operations mainly act as global affine transformations and do not destroy the structural differences between digits.

From a representational standpoint, this construction induces a *finite digit codebook* in the embedding space. Because the digit set is finite, injectivity of this codebook implies the existence of a positive separation margin between different digits, which guarantees robust nearest-neighbor recoverability under small perturbations. This property underlies *digit recoverability* and, by extension, *numeracy preservation* under fixed precision, since each digit can be recovered independently from its embedding.

The use of multiple wavelets and multiple scales further enhances this separation. Concatenating coefficients across wavelet-scale pairs cannot decrease pairwise distances between digit embeddings and typically increases them, thereby improving or maintaining the separation margin. This explains why TempoWave exhibits enhanced discriminability compared to single-scale or single-frequency numeric encodings, as formally analyzed in the appendix.

Crucially, we also analyze how the induced digit codebook behaves under common normalization layers in Transformers. We show that LayerNorm and RMSNorm can only collapse two embeddings under highly restricted affine conditions. As long as the normalized digit codebook remains injective, digit identities remain uniquely recoverable after normalization. Empirically, the multi-wavelet construction yields well-separated digit embeddings that remain distinct throughout the LLM.

Full-context prompt example

Input: Historical load time series (48 half-hour points), e.g., "..., -0.3849, -0.4859, -0.6162, -0.7185, ..."
Context: Region: VIC; Dates: 2021-05-12 → 2021-05-13 (weekday, non-holiday); Resolution: 30 minutes; Horizon: next 24 hours (48 points); Auxiliary features: daily weather statistics (temperature, humidity, pressure).
Instruction: Predict the next 48 load values using the time series and contextual information. Similarity is assessed using Catch22 statistical descriptors to preserve autocorrelation, periodicity, and fluctuation patterns.
Output: Predicted sequence only, e.g., "..., 0.3918, 0.3817, 0.4148, 0.4327, 0.4201, ..."

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate TempoWave on context-enriched forecasting benchmarks where each time series segment is paired with additional textual or event-based context. First, we use the CGTSF dataset released via Hugging Face Datasets [Wang *et al.*, 2025], which contains three collections: MSPG (solar power generation from 27 sites in Melbourne, 2021–2022, 15-minute frequency), LEU (electricity usage from 16 London households, 2012–2013, 30-minute

	AUL		BIT		MSPG		PTF		LEU	
Model	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Time Series Forecasting Model										
DLinear	0.5955	0.4977	1.5160	1.4051	0.4287	0.2901	0.4632	0.3268	0.6422	0.5355
Autoformer	0.8336	0.6828	1.4385	1.2953	0.4501	0.2998	0.3706	0.2288	0.6805	0.5937
TimesNet	0.5431	0.4595	1.2870	1.1874	0.3821	0.2792	0.3437	0.2238	0.3895	0.3303
PatchTST	0.4885	0.4086	1.2723	1.1684	0.4551	0.3098	0.3755	0.2492	0.4536	0.3634
iTransformer	0.6054	0.4948	1.3875	1.2675	0.3970	0.2711	0.3363	0.2234	0.6049	0.4948
N-BEATS	0.7158	0.5874	1.3603	1.1843	0.4305	0.2891	0.3713	0.2306	0.7480	0.5851
Time Series Foundation Model										
TimeLLM	0.7660	0.6012	1.3127	1.1538	0.4378	0.3267	0.5744	0.4730	0.4925	0.3701
Chronos	0.5970	0.4892	1.3348	1.1355	0.4103	0.2393	0.3606	0.2828	0.4184	0.2441
Moirai	0.5806	0.4854	1.3924	0.9823	0.3760	0.2368	0.2141	0.1681	0.3271	0.2267
ChatTime-Chat	0.3639	0.3050	0.8357	0.7421	0.3123	**0.1917**	**0.1610**	0.1276	**0.1557**	0.1253
Fourier Embedding + LLM										
FoNE-Qwen2.5-1.5b-instruct	0.3649	0.3045	1.7141	1.5202	0.3006	0.2660	0.2915	0.2495	0.2065	0.1378
TempoWAVE Embedding + LLM										
MWNE-Qwen2.5-1.5b-instruct	**0.3391**	**0.2739**	**0.7979**	**0.6978**	**0.2950**	0.1956	0.1706	**0.1212**	0.1681	**0.1095**
Relative ↓ Improvement over Previous Best (%)										
TempoWAVE vs Previous SOTA	**7.3**%	**11.2**%	**4.7**%	**6.3**%	**1.9**%	-2.0%	-5.9%	**5.3**%	-7.9%	**14.4**%

Table 1: **Forecasting performance (RMSE/MAE) across five context-enriched datasets.** TempoWAVE achieves new SOTA on 7/10 metrics and ranks second on the remaining three. Best values are **bolded**, second-best are underlined. The last row reports the relative ↓ improvement of MWNE over the previous best method for each metric.

frequency), and PTF (traffic flow from 32 Paris detectors in Paris, 2012, hourly frequency). Each example includes a historical numerical window and associated context such as background descriptions, weather information (from Open-Meteo), date and holiday indicators, and curated news text when available. We follow the official data splits and preprocessing protocol provided by the dataset source.

We additionally use the context-aware forecasting datasets from [Wang *et al.*, 2024], including Australia (AUL) and Bitcoin (BIT), which pair time series with relevant news articles. For AUL and BIT, we follow the original preprocessing, normalization, and train/validation/test splits to ensure comparability with prior work.

Task formulation. Given a historical window of observations and its associated context, the model predicts the next k future values. We adopt a generative formulation: each numeric value is rendered into a fixed-precision string (e.g., `V.FFFF`) and the model generates future values as token sequences. During fine-tuning, we minimize the standard cross-entropy loss over next-token prediction.

Decoding and numeric parsing. At inference time, generated token sequences are converted back to real values by parsing the fixed-precision numeric strings. An example of the full-context prompt is detailed in the accompanying text box. If a generated output violates the numeric format (e.g., missing digits or containing non-numeric tokens), we apply a deterministic fallback parsing rule; if parsing still fails, the prediction for that step is treated as invalid and is counted in the evaluation according to the protocol. All formatting and parsing rules are fixed across methods to ensure a fair comparison.

Evaluation metrics. Forecasting accuracy is measured using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) across multiple prediction horizons. Metrics are computed after inverting dataset-specific normalization when applicable, following the evaluation protocol of the corresponding benchmarks.

Baselines and fairness. We compare TempoWave-enhanced LLMs against a comprehensive set of baselines. **(i) LLM-based baselines.** These methods use the same prompt templates and contextual inputs as TempoWave and differ only in the numeric interface, including standard tokenization and alternative input adaptation strategies. **(ii) Time-series-specific baselines.** We also report results from established forecasting models that primarily operate on numerical history, including DLinear [Zeng *et al.*, 2023], N-BEATS [Oreshkin *et al.*, 2019], Informer [Zhou *et al.*, 2021], Autoformer [Wu *et al.*, 2021], and TimesNet [Wu *et al.*, 2023], as well as large-scale time series foundation models such as Chronos [Ansari *et al.*, 2024] and Moirai [Woo *et al.*, 2024]. **(iii) Context-aware and embedding-interface baselines.** We include ChatTime [Wang *et al.*, 2025] as a representative multimodal LLM system for forecasting with text context, and FoNE [Zhou *et al.*, 2025] as an alternative

	AUL		BIT		MSPG		PTF		LEU	
Setting	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
w/o context	0.3809	0.3149	0.8356	0.7261	0.3218	0.1917	0.2981	0.2391	0.2099	0.1613
w/o Catch22	0.3759	0.3121	0.8205	0.7102	0.3023	0.2057	0.2647	0.2119	0.1941	0.1266
w/o Situational context	0.3482	0.2823	0.8044	0.7007	0.3123	**0.1901**	0.1795	0.1272	0.1843	0.1284
Full context	**0.3391**	**0.2739**	**0.7979**	**0.6978**	**0.2950**	0.1956	**0.1706**	**0.1212**	**0.1681**	**0.1095**

Table 2: Ablation Study: Forecasting performance (RMSE/MAE) across datasets under different context settings.

numerical embedding interface.

4.2 Main Results

Table 1 summarizes forecasting performance on five context-enriched benchmarks spanning news-driven series (AUL, BIT) and sensor or infrastructure series (MSPG, PTF, LEU). Overall, TempoWave establishes a new state of the art on 7 out of 10 reported metrics and achieves top-2 performance on all metrics. Relative to the previous best method per metric (last row of Table 1), TempoWave yields an average 7.0% relative improvement on MAE across datasets, with the largest gains on LEU (14.4%) and AUL (11.2%).

Dataset-wise improvements and robustness. TempoWave delivers the most consistent gains on news-driven datasets. On AUL, TempoWave improves both RMSE and MAE over the previous best by 7.3% and 11.2%, respectively. On BIT, TempoWave achieves 4.7% (RMSE) and 6.3% (MAE) improvements over the previous best. On MSPG, TempoWave achieves the best RMSE (1.9% improvement) while remaining close to the best MAE (within 2.0% relative to the previous best). For PTF and LEU, TempoWave attains the best MAE (5.3% and 14.4% improvements), and achieves the second-best RMSE with small absolute gaps to the best baseline (0.0096 on PTF, 0.0124 on LEU).

MAE improves more consistently than RMSE. A recurring pattern is that TempoWave improves MAE more consistently than RMSE. In particular, TempoWave reduces MAE on 4/5 datasets, while RMSE improvements are observed on 3/5 datasets. This discrepancy is expected because RMSE emphasizes rare large deviations, whereas MAE better reflects typical per-step errors.

Comparison to numeric interfaces and time-series baselines. Compared with the Fourier-based numeric interface (FoNE) using the same LLM backbone, TempoWave is substantially more robust across domains. The advantage is most prominent on BIT, where TempoWave reduces RMSE from 1.71 to 0.80 and MAE from 1.52 to 0.70, indicating improved generalization under highly non-stationary and event-driven dynamics. Finally, TempoWave-enhanced LLMs outperform classic time-series forecasting models across all datasets in Table 1, highlighting the benefit of combining external context with a numerically grounded embedding interface.

5 Analysis

5.1 Ablation Study on Contextual Information

To better understand how TempoWave interacts with different forms of contextual information in time series forecasting, we conduct a systematic ablation study over four context configurations, as summarized in Table 2. Across five diverse datasets (AUL, BIT, MSPG, PTF, and LEU), we progressively remove components from the full context setting to isolate their individual and combined effects.

Overall impact of contextual information. The results show a clear and consistent trend: incorporating richer contextual information leads to improved forecasting performance across all datasets. The full context setting achieves the best RMSE on all five datasets and the best MAE on four out of five datasets. In contrast, removing all contextual information results in the weakest performance, indicating that TempoWave alone, while effective, benefits substantially from complementary contextual signals. This trend is particularly pronounced on news-driven datasets such as AUL and BIT, where RMSE improves from 0.3809 to 0.3391 on AUL and from 0.8356 to 0.7979 on BIT when moving from no context to full context.

Contribution of different context components. Comparing partial ablations reveals that different types of context contribute in distinct and complementary ways. Removing Catch22 features (*w/o Catch22*) leads to noticeable degradation across most datasets, suggesting that statistical descriptors capturing autocorrelation, periodicity, and distributional properties provide strong global signals for forecasting. Indeed, the *w/o Catch22* setting consistently underperforms the full context configuration. Conversely, removing situational context (*w/o situational context*) primarily affects datasets with strong external dependencies, such as AUL and BIT, where date and domain-related information play a more prominent role.

Dataset-specific behavior. The relative importance of context components varies across datasets. For infrastructure and sensor-driven datasets (MSPG, PTF, and LEU), Catch22 features alone already provide strong performance, in some cases matching or approaching the full context results. For example, on MSPG, the *w/o situational context* setting achieves the best MAE (0.1901), indicating that short-term statistical regularities dominate forecasting performance. In contrast, on AUL and BIT, which are influenced by external events and news, the full context setting yields the largest

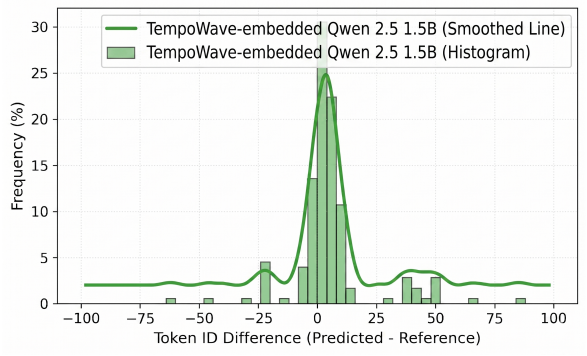

Figure 2: Token ID difference distribution between predicted tokens and their reference counterparts for **TempoWave-embedded Qwen 2.5 1.5B model**, under the top-10 prediction setting. The histogram illustrates raw frequency, while the smoothed curve highlights the overall trend. The sharp concentration around zero indicates strong local proximity in token prediction.

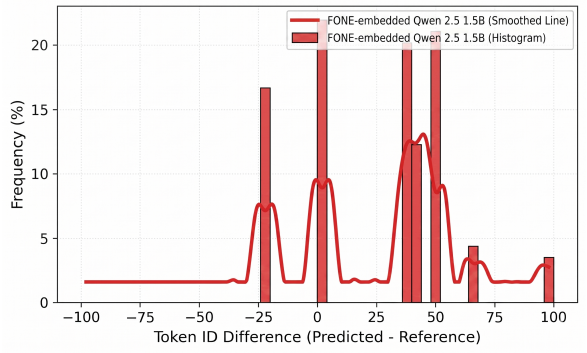

Figure 3: Token ID difference distribution between predicted tokens and their reference counterparts for **FoNE-embedded Qwen 2.5 1.5B model** (baseline), under the top-10 prediction setting. The histogram illustrates raw frequency, while the smoothed curve highlights the overall trend. The sharp concentration around zero indicates strong local proximity in token prediction.

gains, highlighting the importance of integrating situational and textual information with TempoWave.

5.2 Embedding Alignment via Next Token Proximity

To evaluate the semantic and structural alignment of different embedding strategies, we analyze the distribution of token ID proximity between the model's predicted next token and the immediately preceding token in the input prompt. This probing task is particularly informative in our setting, where tokens represent numerical values derived from time series data. A well-structured embedding should induce a smooth, symmetric distribution reflecting temporal continuity. Our method, as shown in Figure 2, exhibits a clear unimodal, approximately Gaussian distribution centered around zero, indicating that the model learns to predict numerically coherent tokens aligned with the underlying time series dynamics. In contrast, the FoNE baseline in Figure 3, evaluated with the same backbone but without an explicit numeric inductive

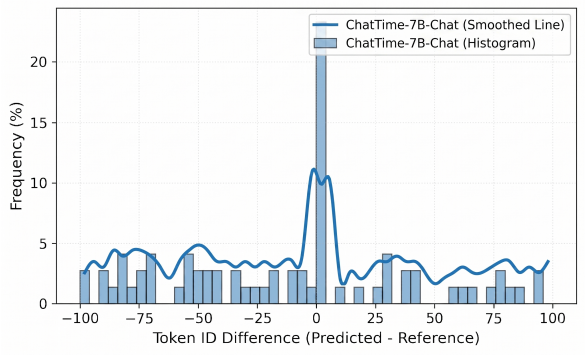

Figure 4: Token ID difference distribution between predicted tokens and their reference counterparts for **ChatTime-7B-Chat model** (baseline), under the top-10 prediction setting. The histogram illustrates raw frequency, while the smoothed curve highlights the overall trend. The sharp concentration around zero indicates strong local proximity in token prediction.

bias, exhibits a flatter and more irregular distribution, indicating weaker alignment with underlying numeric trends. More notably, a standard pretrained baseline without our embedding augmentation in Figure 4 exhibits a sharp, anomalous spike in one bin, revealing a tendency to overfit by repeatedly predicting a fixed token, regardless of local context. These results underscore the effectiveness of our embedding approach in capturing latent numerical semantics and encoding smooth transitions that mirror real-world time series behavior.

6 Conclusion

While directly applying large language models (LLMs) to time series analysis remains challenging due to the mismatch between continuous values and discrete token interfaces, the potential payoff is substantial. In this paper, we proposed Multi-Wavelet Number Embedding (TempoWave), a numerically grounded embedding interface that leverages multi-resolution wavelet features to bridge the numerical–textual modality gap for time series forecasting. Extensive experiments on five diverse benchmarks show that TempoWave consistently improves LLM-based forecasters, outperforming strong specialized time series models and alternative numeric embedding approaches in most settings. Empirically, TempoWave is more robust under non-stationarity and extreme values, and exhibits favorable optimization behavior, including smoother training dynamics, resilience to digit-level perturbations, and stable interaction with common normalization layers. Ablation results further highlight that contextual information is complementary to TempoWave and contributes to the strongest overall performance. Together, these findings advance LLM-based forecasting by coupling LLMs' contextual reasoning with a more faithful numeric interface. A promising direction for future work is to investigate whether TempoWave also benefits non-contextual forecasting pipelines that rely on discretization or binning-based tokenization of time series values.

Ethical Statement

There are no ethical issues.

Acknowledgements

This work is partially supported by the NSF Award #2425919, and NSF Award #2413417. The funding from these sources has been a cornerstone in enabling us to bring our project to fruition. We are also deeply grateful to the anonymous reviewers for their rigorous review process. Their detailed comments and constructive suggestions have significantly contributed to the improvement of this paper.

References

[Ansari *et al.*, 2024] 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.

[Cao *et al.*, 2020] Defu Cao, Yujing Wang, Juanyong Duan, Ce Zhang, Xia Zhu, Congrui Huang, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, et al. Spectral temporal graph neural network for multivariate time-series forecasting. *Advances in neural information processing systems*, 33:17766–17778, 2020.

[Cao *et al.*, 2021] Defu Cao, Jiachen Li, Hengbo Ma, and Masayoshi Tomizuka. Spectral temporal graph neural network for trajectory prediction. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1839–1845, 2021.

[Cao *et al.*, 2023a] Defu Cao, James Enouen, Yujing Wang, Xiangchen Song, Chuizheng Meng, Hao Niu, and Yan Liu. Estimating treatment effects from irregular time series observations with hidden confounders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6897–6905, 2023.

[Cao *et al.*, 2023b] Defu Cao, Yixiang Zheng, Parisa Hassanzadeh, Simran Lamba, Xiaomo Liu, and Yan Liu. Large scale financial time series forecasting with multi-faceted model. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, ICAIF '23, page 472–480, New York, NY, USA, 2023. Association for Computing Machinery.

[Cao *et al.*, 2024a] 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. In *The Twelfth International Conference on Learning Representations*, 2024.

[Cao *et al.*, 2024b] Defu Cao, Wen Ye, Yizhou Zhang, and Yan Liu. Timedit: General-purpose diffusion transformers for time series foundation model. *arXiv preprint arXiv:2409.02322*, 2024.

[Cao *et al.*, 2025] Defu Cao, Michael Gee, Jinbo Liu, Hengxuan Wang, Wei Yang, Rui Wang, and Yan Liu. Conversational time series foundation models: Towards explainable and effective forecasting. *arXiv preprint arXiv:2512.16022*, 2025.

[Cao *et al.*, 2026] Defu Cao, Wen Ye, Yizhou Zhang, Sam Griesemer, and Yan Liu. PINFDit: Energy-based physics-informed diffusion transformers for general-purpose time series tasks. In *The Fourteenth International Conference on Learning Representations*, 2026.

[Das *et al.*, 2024] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*, 2024.

[Gillman *et al.*, 2025] Nate Gillman, Daksh Aggarwal, Michael Freeman, and Chen Sun. Fourier head: Helping large language models learn complex probability distributions. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Goswami *et al.*, 2024] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. In *International Conference on Machine Learning*, pages 16115–16152. PMLR, 2024.

[Gruver *et al.*, 2024] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36, 2024.

[Hu *et al.*, 2025] Yuxiao Hu, Qian Li, Dongxiao Zhang, Jinyue Yan, and Yuntian Chen. Context-alignment: Activating and enhancing LLMs capabilities in time series. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Jia *et al.*, 2024] Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. Gpt4mts: Prompt-based large language model for multimodal time-series forecasting. In *The 14th Symposium on Educational Advances in Artificial Intelligence (EAAI-24)*, 2024.

[Jin *et al.*, 2024] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[Li *et al.*, 2026] Jiate Li, Defu Cao, Li Li, Wei Yang, Yuehan Qin, Chenxiao Yu, Tiannuo Yang, Ryan A Rossi, Yan Liu, Xiyang Hu, et al. "someone hid it!": Query-agnostic black-box attacks on LLM-based retrieval. In *Forty-third International Conference on Machine Learning*, 2026.

[Liu *et al.*, 2022] Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *Advances in Neural Information Processing Systems*, 2022.

[Lubba *et al.*, 2019] Carl H Lubba, Sarab S Sethi, Philip Knaute, Simon R Schultz, Ben D Fulcher, and Nick S Jones. catch22: Canonical time-series characteristics: Selected through highly comparative time-series analysis. *Data mining and knowledge discovery*, 33(6):1821–1852, 2019.

[Merrill *et al.*, 2024] Mike A Merrill, Mingtian Tan, Vinayak Gupta, Thomas Hartvigsen, and Tim Althoff. Language models still struggle to zero-shot reason about time series. In *EMNLP (Findings)*, 2024.

[Nie *et al.*, 2023] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR '23)*, 2023.

[OpenAI, 2023] OpenAI. Gpt-4 technical report, 2023.

[Oreshkin *et al.*, 2019] Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437*, 2019.

[Talukder *et al.*, 2024] Sabera J Talukder, Yisong Yue, and Georgia Gkioxari. Totem: Tokenized time series embeddings for general time series analysis. *Transactions on Machine Learning Research*, 2024.

[Wang *et al.*, 2024] Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection. In *Neural Information Processing Systems*, 2024.

[Wang *et al.*, 2025] Chengsen Wang, Qi Qi, Jingyu Wang, Haifeng Sun, Zirui Zhuang, Jinming Wu, Lei Zhang, and Jianxin Liao. Chattime: A unified multimodal time series foundation model bridging numerical and textual data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 12694–12702, 2025.

[Wang *et al.*, 2026a] Xiaoda Wang, Ching Chang, Defu Cao, Kaiqiao Han, Fang Sun, Yue Huang, Minxiao Wang, Chang Xu, Xiao Luo, Runze Yan, et al. Position: Beyond prediction: Toward verifiable physiological waveform reasoning with foundation models and agentic LLMs. In *Forty-third International Conference on Machine Learning Position Paper Track*, 2026.

[Wang *et al.*, 2026b] Xiaoda Wang, Kaiqiao Han, Yuhao Xu, Xiao Luo, Yizhou Sun, Wei Wang, and Carl Yang. Se-diff: Simulator and experience enhanced diffusion model for comprehensive ecg generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

[Weng *et al.*, 2026] Muyan Weng, Defu Cao, Wei Yang, Yashaswi Sharma, and Yan Liu. Temporalbench: A benchmark for evaluating llm-based agents on contextual and event-informed time series tasks. *arXiv preprint arXiv:2602.13272*, 2026.

[Woo *et al.*, 2024] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Forty-first International Conference on Machine Learning*, 2024.

[Wu *et al.*, 2021] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 101–112, 2021.

[Wu *et al.*, 2023] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *The Eleventh International Conference on Learning Representations*, 2023.

[Yang *et al.*, 2025a] Wei Yang, Defu Cao, and Yan Liu. Foundation models for demand forecasting via dual-strategy ensembling. *arXiv preprint arXiv:2507.22053*, 2025.

[Yang *et al.*, 2025b] Wei Yang, Muyan Weng, Jiacheng Pang, Defu Cao, Heng Ping, Peiyu Zhang, Shixuan Li, Yue Zhao, Qiang Yang, Mengdi Wang, et al. Toward evolutionary intelligence: Llm-based agentic systems with multi-agent reinforcement learning. *Available at SSRN 5819182*, 2025.

[Yang *et al.*, 2026] Wei Yang, Defu Cao, Jiacheng Pang, Muyan Weng, and Yan Liu. Adaptive collaboration with humans: Metacognitive policy optimization for multi-agent LLMs with continual learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

[Ye *et al.*, 2025] Wen Ye, Jinbo Liu, Defu Cao, Wei Yang, and Yan Liu. When llm meets time series: Can llms perform multi-step time series reasoning and inference. *arXiv preprint arXiv:2509.01822*, 2025.

[Ye *et al.*, 2026] Wen Ye, Wei Yang, Defu Cao, Yizhou Zhang, Lumingyuan Tang, Jie Cai, and Yan Liu. TS-reasoner: Domain-oriented time series inference agents for reasoning and automated analysis. *Transactions on Machine Learning Research*, 2026.

[Zeng *et al.*, 2023] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128, 2023.

[Zhang *et al.*, 2022] Yizhou Zhang, Defu Cao, and Yan Liu. Counterfactual neural temporal point process for estimating causal influence of misinformation on social media. *Advances in Neural Information Processing Systems*, 35:10643–10655, 2022.

[Zhang *et al.*, 2024] Yizhou Zhang, Lun Du, Defu Cao, Qiang Fu, and Yan Liu. Guiding large language models with divide-and-conquer program for discerning problem solving. *arXiv preprint arXiv:2402.05359*, 2024.

[Zhou and Yu, 2025] Zihao Zhou and Rose Yu. Can LLMs understand time series anomalies? In *The Thirteenth International Conference on Learning Representations*, 2025.

[Zhou *et al.*, 2021] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of AAAI*, 2021.

[Zhou *et al.*, 2025] Tianyi Zhou, Deqing Fu, Mahdi Soltanolkotabi, Robin Jia, and Vatsal Sharan. Fone: Precise single-token number embeddings via fourier features. *arXiv preprint arXiv:2502.09741*, 2025.