Spectral Predictability as a Fast Reliability Indicator for Time Series Forecasting Model Selection

Oliver Wang, Pengrui Quan, Kang Yang, Mani Srivastava*

Electrical and Computer Engineering
University of California, Los Angeles

Abstract

Practitioners deploying time series forecasting models face a dilemma: exhaustively validating dozens of models is computationally prohibitive, yet choosing the wrong model risks poor performance. We show that *spectral predictability* Ω—a simple signal processing metric—systematically stratifies model family performance, enabling fast model selection. We conduct controlled experiments in four different domains, then further expand our analysis to 51 models and 28 datasets from the GIFT-Eval benchmark. We find that large time series foundation models (TSFMs) systematically outperform lightweight task-trained baselines when Ω is high, while their advantage vanishes as Ω drops. Computing Ω takes seconds per dataset, enabling practitioners to quickly assess whether their data suits TSFM approaches or whether simpler, cheaper models suffice. We demonstrate that Ω stratifies model performance, offering a practical first-pass filter that reduces validation costs while highlighting the need for models that excel on genuinely difficult (low-Ω) problems rather than merely optimizing easy ones. Code and data available at https://github.com/nesl/Spectral-Predictability-TS/.

Introduction

Large time series foundation models (TSFMs) for time series forecasting promise broad improvements by leveraging massive pretraining (Ye et al. 2024; Li et al. 2025; Liang et al. 2024; Ansari et al. 2024; Gruver et al. 2024). Yet empirical evidence remains mixed; simple baselines such as DLinear often match or surpass complex architectures (Tan et al. 2024; Zeng et al. 2022; Li et al. 2025). Practitioners face a practical challenge: *how to choose which model to deploy without exhaustively validating every option?*

Comprehensive validation is impractical. Consider a practitioner with a dozen or more candidate models and a new dataset: training and validating all models requires substantial compute, time, and engineering effort. Worse, this process provides no insight into *why* certain models work better, making it difficult to generalize lessons to future datasets.

We propose *spectral predictability* Ω—a simple, fast-to-compute signal property—as a reliability indicator that narrows the model search space *before* expensive validation begins. Grounded in signal processing, Ω quantifies the concentration of a series' power spectrum: high Ω reflects structured, repeatable patterns; low Ω indicates diffuse, irregular signals. Computing Ω takes seconds on a commodity device, yet we show it systematically stratifies model performance.

Our key finding. Large zero-shot[1] models, applied without fine-tuning, show consistent advantages in high-Ω regimes across diverse domains. Practitioners can compute Ω to determine whether zero-shot or lightweight models are likely to perform best, reducing validation cost. As Ω decreases, model performance converges, underscoring the need for methods that better handle difficult (low-Ω) data.

In summary, this paper makes three contributions:

- **A fast, interpretable forecastability indicator.** We introduce spectral predictability Ω as a dataset-level measure of frequency-domain concentration that can be computed in seconds without model training, providing a lightweight proxy for time-series forecastability.
- **Evidence that Ω stratifies intrinsic difficulty.** Across controlled experiments on synthetic and real-world datasets (CarbonCast, PEMS, Fitbit), forecasting error decreases monotonically with increasing Ω, indicating that Ω captures an important axis of inherent predictability.
- **Large-scale model-family behavior as a function of Ω.** On 28 GIFT-Eval datasets and 51 forecasting models, we find that zero-shot TSFMs achieve their largest gains in high-Ω regimes (up to 60% over statistical and deep-learning baselines), while this advantage diminishes in low-Ω settings.

Implications.

- **Practical deployment: shortlist models before validation.** Because Ω is fast to compute, it can serve as a first-pass screening signal to narrow the model search space and reduce validation cost, especially when evaluating many candidate forecasters is expensive.

[*] Mani Srivastava holds concurrent appointments as a Professor of ECE and CS (joint) at the University of California, Los Angeles, and as an Amazon Scholar at Amazon. This paper describes work performed at UCLA and is not associated with Amazon.

[1] We use GIFT-Eval's model taxonomy where "zero-shot" refers to TSFMs deployed with their original pretrained weights. See Large-Scale Analysis Results for full definitions.

- **A clear failure mode: low-Ω as an open frontier.** The low-Ω regime consistently corresponds to settings where all model families struggle, suggesting that irregular or weakly periodic signals remain a key bottleneck and motivating the development of methods that are robust beyond strong spectral structure.

This work directly aligns with the AI4TS goal of advancing both the theoretical and practical aspects of time series analysis. By connecting classical signal-processing theory with modern foundation models, our framework offers an interpretable and computationally efficient approach to *forecasting, interpretable model selection, and reliability estimation*—core topics of AI4TS. The study bridges domains such as IoT, energy, healthcare, and mobility, illustrating how spectral measures can help unify representation, prediction, and evaluation in real-world time series applications.

Related Work

Simplicity *vs.* Capacity. Despite scaling trends (Shi et al. 2024, 2025), lightweight baselines remain competitive (Zeng et al. 2022; Miller et al. 2024). Comparative studies (Goswami et al. 2024; Jin et al. 2024) rarely explain *why* performance varies across domains, leaving practitioners without guidance for model selection.

LLMs for Time Series. Methods include direct tokenization, architectural adaptation, and adapter-based fine-tuning (Gruver et al. 2024; Ansari et al. 2024). Ablations question how much LLM pretraining contributes (Tan et al. 2024; Jin et al. 2024; Elsayed et al. 2021). We build on Jin et al. (2024) for our codebase and initial experiments are based on variations on their LLAMA-7B backbone structure, which will be explained further in the Controlled Experiment Results.

Forecastability and Reliability. Forecastability metrics such as spectral entropy, approximate entropy, and seasonality strength relate to signal difficulty (Tang et al. 2024; Wu et al. 2023; Wang, Klee, and Roos 2025; Guntu et al. 2020). While these metrics characterize data properties, they have not been systematically used to guide model selection at deployment time.

Our contribution is not the Ω metric itself—spectral entropy is well-established—but rather the empirical discovery that zero-shot models exhibit a unique, systematic relationship with Ω that other model families do not. This differential response enables targeted model selection: for high-Ω data, the choice is clear; for low-Ω data, the advantages of these large models disappear. Table 1 contrasts our approach with existing alternatives.

Table 1: Comparison of model selection approaches. Ω uniquely provides model-family-specific guidance with minimal computation.

Approach	Speed	Model Guidance	Interpretable
Spectral entropy (Wang 2025)	Fast	×	✓
Approx. entropy (Pincus 1991)	Fast	×	×
Validation subset	Medium	✓	×
Meta-learning (Talagal 2024)	Slow	✓	×
AutoML (Salehin 2024)	Very slow	✓	×
Spectral Predictability Ω (ours)	**Fast**	✓	✓

Model Selection. Traditional model selection requires training and validating multiple candidates, which is resource-intensive. Meta-learning and AutoML approaches attempt to automate this process but still require significant computation (Li et al. 2025). Our approach complements these methods by providing a fast preliminary filter based on data properties alone, enabling practitioners to focus expensive validation on a smaller subset of promising models.

Spectral Predictability Ω

We quantify the inherent forecastability of a time series using spectral predictability Ω, a metric grounded in information theory and signal processing. Ω captures how concentrated the energy is in the frequency domain: periodic series with strong seasonal patterns have concentrated spectra and high predictability, while noisy or irregular series yield diffuse spectra and low predictability (Wang, Klee, and Roos 2025; Guntu et al. 2020).

Let $\{x_t\}_{t=1}^{T}$ be a univariate series of length T. Apply a Hann taper and remove the DC component, then compute the FFT. Define the one-sided power spectral density:

$$P_k = |\hat{x}_k|^2, \quad k = 1, \dots, K, \quad K = \lfloor T/2 \rfloor,$$

where $\hat{x}_k$ denotes the k-th frequency component (DC excluded). Normalize to obtain a probability distribution $p_k = P_k / \sum_{j=1}^{K} P_j$ and compute spectral entropy:

$$H(x) = -\sum_{k=1}^{K} p_k \log p_k.$$

Spectral predictability is defined by normalizing entropy by its maximum $H_{\max} = \log K$:

$$\Omega(x) = 1 - \frac{H(x)}{H_{\max}}, \quad \Omega \in [0, 1].$$

High Ω indicates concentrated spectra (more predictable); low Ω indicates diffuse spectra (less predictable).

Computational Efficiency. Computing Ω requires only a single FFT pass, taking seconds on a standard laptop for typical forecasting datasets (thousands to millions of time points)—orders of magnitude faster than training even a single model. This makes Ω a practical preprocessing step for model selection.

Experimental Overview

We assess spectral predictability (Ω) through two stages:

- **Controlled Experiments:** Synthetic signals with tunable Ω and three real datasets (CarbonCast, PEMS, Fitbit) test how forecasting error changes with Ω. Models include TimeLLM (with both LLAMA3.2-1B and GPT2-130M backbones), randomly initialized backbone, and DLinear, evaluated by sMAPE and MSE.
- **Large-Scale Analysis:** Using 51 models and 28 datasets from GIFT-Eval, we compute dataset-level Ω to compare statistical, deep-learning, pretrained, and zero-shot models across predictability levels.

These experiments reveal how Ω captures forecasting difficulty and guides model selection.

Controlled Experiment Results: Establishing the Effect of Ω

To test whether spectral predictability genuinely affects forecasting difficulty—and can be systematically manipulated—we designed controlled experiments across four domains with varying characteristics:

Synthetic Data. We created synthetic Fourier signals explicitly engineered to span Ω values from 0.2 to 0.8. By controlling the spectral entropy directly through the frequency components, we generated time series with predetermined predictability levels.

Real-World Domains. We also tested three diverse real-world datasets: (i) CarbonCast: hourly energy generation (Maji, Shenoy, and Sitaraman 2022); (ii) PEMS: hourly traffic flow (Wang et al. 2024); and (iii) Fitbit: minute-level heart rate (Furberg et al. 2016). These domains exhibit natural variation in Ω arising from different underlying processes, allowing us to verify that patterns observed in synthetic data generalize to realistic conditions.

Models. We evaluated four representative architectures: (i) TimeLLM pretrained with frozen Llama3.2-1B weights; (ii) the same architecture with random initialization; (iii) GPT2-130M; and (iv) DLinear (Zeng et al. 2022; Radford et al. 2019). All models used 512-step context and 96-step forecast horizon. Error was measured by the Symmetric Mean Absolute Percentage Error (sMAPE). sMAPE is a scale-normalized accuracy metric that lies in $[0, 2]$ and is defined for a forecast $\hat{y}_t$ of target y_t over T timesteps as

$$\text{sMAPE} = \frac{1}{T}\sum_{t=1}^{T}\frac{2|\hat{y}_t - y_t|}{|\hat{y}_t| + |y_t|}.$$

Lower values indicate better predictive accuracy, and because the denominator rescales by the magnitude of both the forecast and the ground truth at each timestep, sMAPE is comparable across datasets with different units and scales.

Further training details are in the Appendix.

Consistency Across Metrics. Our primary analysis uses sMAPE because it is scale-normalized, enabling meaningful aggregation and comparison across datasets with widely different magnitudes and units. This is important in our setting, where we evaluate many heterogeneous domains. However, we verify the robustness of our observations using the popular MSE metric on controlled experiments. Table 2 shows the relationship between MSE and Ω exhibits consistent negative correlations across all domains (Pearson r ranging from -0.377 to -0.750), confirming that the core pattern—error decreases as predictability increases—holds across error metrics. The consistency between sMAPE and MSE results suggests our findings are not artifacts of metric choice, though future work should examine probabilistic scores (CRPS, interval coverage) for additional validation.

Key Findings. Fig. 1 shows noticeable patterns across all domains: forecasting error systematically decreases as Ω increases. This effect is most pronounced in Synthetic, where we engineered Ω directly, providing strong evidence that spectral predictability correlates with difficulty. The pattern replicates in CarbonCast (energy) and, to a lesser extent, in PEMS (traffic) and Fitbit (wearables).

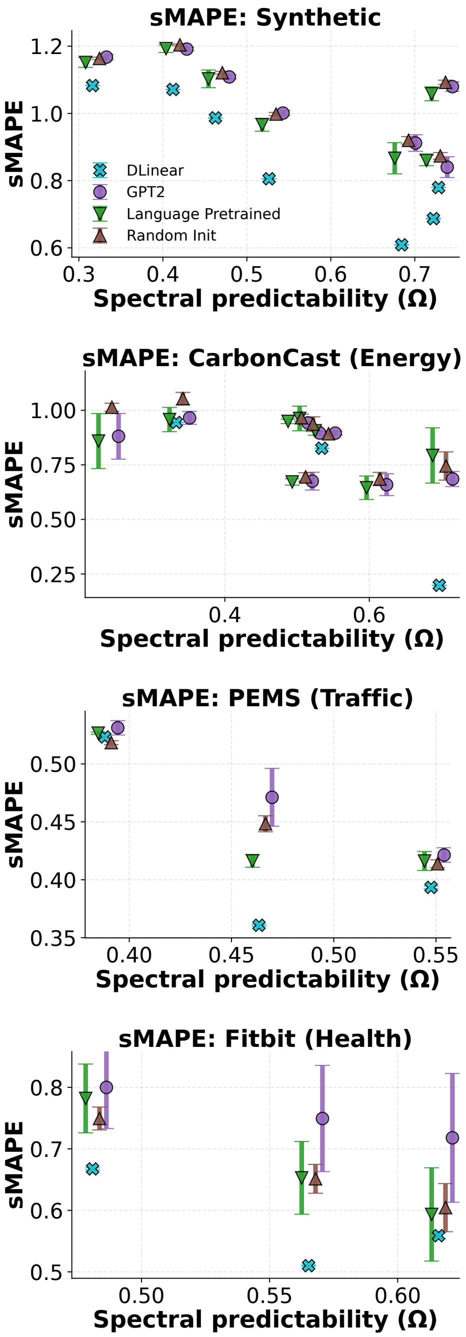

Figure 1: **Spectral predictability systematically affects forecasting difficulty.** Across synthetic and real-world domains, sMAPE declines as Ω increases. Error bars show 95% CIs across series. The clearest pattern emerges in synthetic data where Ω is directly controlled. Note that less data was available for PEMS and Fitbit, leading to sparser graphs. Also note that model classes have been slightly offset horizontally for visual clarity.

In Synthetic and CarbonCast, where spectral structure dominates signal characteristics, the Ω-error relationship is nearly monotonic. Models tend to show improved performance at high Ω, with error reductions of 20–40% when

Table 2: Aggregate relationship between MSE and Ω across controlled experiments. Negative correlations indicate that forecasting error decreases as predictability increases, supporting Ω as a proxy for difficulty.

Dataset	Pearson r	Spearman ρ
Synthetic	-0.720	-0.678
CarbonCast	-0.676	-0.740
PEMS	-0.750	-0.708
Fitbit	-0.377	-0.367

moving from $\Omega = 0.3$ to $\Omega = 0.7$.

The effect is weaker in PEMS and Fitbit, likely because other factors—missingness patterns (Fitbit users removing devices), noise characteristics, and domain-specific irregularities—contribute substantially to difficulty beyond spectral properties alone. This suggests Ω is a useful but not exhaustive indicator; practitioners should consider it alongside domain knowledge.

These controlled experiments suggest a key result: spectral predictability systematically stratifies forecasting difficulty. However, these experiments lack scale and leave open a critical question for practitioners: **do different model families—statistical, deep learning, pretrained or TSFM—respond differently to Ω?** Understanding this would enable targeted model selection based on dataset properties. We investigate this next in a more comprehensive setting.

Large-Scale Analysis Results: Model-Family-Specific Responses to Ω

To examine whether different model families exhibit distinct relationships with spectral predictability, we analyzed 51 models from the **GIFT-Eval Time Series Forecasting Leaderboard**, spanning statistical (AutoETS/Theta/ARIMA), deep learning (PatchTST, iTransformer), pretrained TSFMs (Chronos, Lag Llama), and zero shot TSFMs (Moirai, TimesFM 2.5); full list in Appendix. We collected their reported sMAPE performance across 28 datasets spanning energy, healthcare, finance, and natural domains (Aksu et al. 2024). Each model was categorized as *statistical*, *deep-learning*, *pretrained*, or *zero-shot* following GIFT-Eval's taxonomy. Further model type categories include *fine-tuned* and *agentic*, though they are not the focus of this study due to the small number of representatives at the time of writing. All models used and their respective categories are reported in the Appendix. In this context, both pretrained and zero-shot models are large TSFM models applied directly without fine-tuning. However, certain models (eg. TimesFM) were originally trained with some amount of data that is in the GIFT-Eval evaluation dataset. To prevent leakage, these models were then *pretrained* on a leak-free dataset designed by (Aksu et al. 2024) and are considered "pretrained". On the other hand, models labeled as "zero-shot" (such as TimesFM-2.5) are TSFMs with no data leakage in their published weights and thus not "re-pretrained".

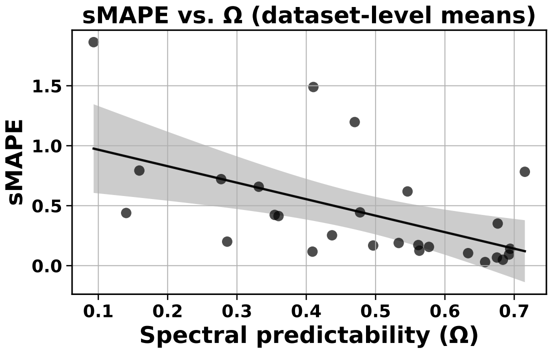

Figure 2: **Predictability-error relationship at scale.** Across 28 datasets and 51 models, average error (sMAPE) declines with increasing spectral predictability Ω. Each point represents an average of all models on one dataset. We fit an ordinary least squares line of best fit with 95% confidence interval for visualization.

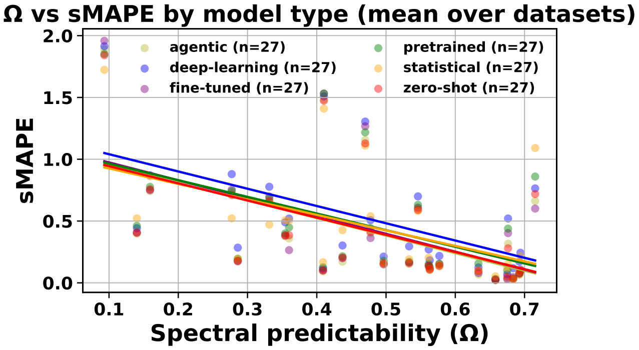

Figure 3: **Predictability-error relationship with model types split out.** The model type classes were taken from GIFT-Eval's classification (Aksu et al. 2024).

Train–test splits were not public, so we computed Ω over each full dataset to characterize its overall spectral properties. This aggregate Ω serves as a dataset-level descriptor that does not inform individual predictions. Our goal was to identify systematic patterns in how different model families respond to varying levels of predictability.

Overall Pattern. Across the 28 datasets, we found a statistically significant monotonic relationship between predictability and error (Spearman $\rho = -0.65$, $p = 1.9 \times 10^{-21}$), confirming that the pattern observed in controlled experiments generalizes at scale (Fig. 2). Results for sMAPE versus Ω by model type are presented in Fig. 3, and as binned averages in Fig. 4, suggesting that this trend is consistent for different model types. To produce the bins in Fig. 4, datasets are grouped into 6 quantile bins of Ω, which ensures that each bin contains a similar number of datasets, preventing high-density Ω regions from dominating the analysis. Each plotted point represents the averaged sMAPE for one of the 4 given model classes in a given Ω regime, with vertical error bars showing the uncertainty across datasets. Model classes are slightly offset horizontally within each bin to avoid overlap and improve visual separation.

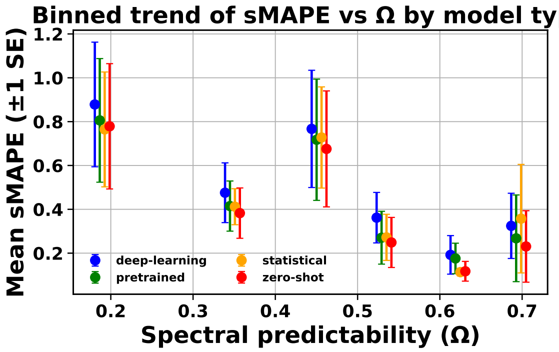

Figure 4: **Model Types Binned for Clarity.** Only model types with more than 4 representative models were chosen to be represented here for robustness and visual clarity.

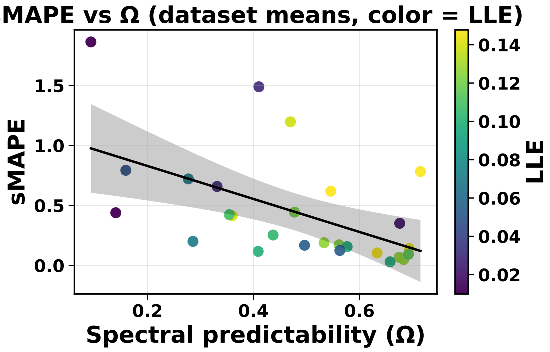

Figure 5: **Relationship with LLE.** A heatmap showing LLE and Omega. A higher LLE corresponds to larger amounts of chaos in a sequence. This plot implies that some of the datasets which exhibit high Ω also have high chaos, which could explain why the sMAPE unexpectedly worsens.

Relationship with Chaos (Largest Lyapunov Exponent)

To investigate whether spectral predictability correlates with chaotic dynamics, we computed the Largest Lyapunov Exponent (LLE) for each dataset. The LLE measures a system's sensitivity to initial conditions: it quantifies the average exponential rate at which two nearby trajectories in the reconstructed state space diverge. Formally,

$$\lambda_{\max} = \lim_{\Delta t \to \infty} \frac{1}{\Delta t} \left\langle \ln \frac{\|\delta \mathbf{x}(t+\Delta t)\|}{\|\delta \mathbf{x}(t)\|} \right\rangle,$$

where $\delta\mathbf{x}(t)$ is an infinitesimal perturbation between two initially close states of the same sequence. Higher LLE indicates more chaotic, less locally predictable dynamics.

Fig. 5 shows a counterintuitive pattern: datasets with higher Ω (more predictable spectra) sometimes exhibit higher LLE values (suggesting more chaos). This apparent paradox arises because spectral predictability and dynamical chaos measure different aspects of time series structure. Ω captures frequency-domain regularity (periodic or quasi-periodic patterns), while LLE measures sensitivity to initial conditions in phase space. A series can have highly structured spectral content (high Ω) while still being chaotic in the deterministic sense. Importantly, this complexity indicates that while higher Ω associates with lower forecasting error, other qualities of the dataset can also have an impact and deserve further investigation.

Model-Family-Specific Patterns. To examine differential responses, we compared relative accuracies between model types for each dataset. For each model pair $A \to B$ evaluated on dataset i:

$$\Delta^{\text{sMAPE}}_{A\to B}(i) = 100 \times \frac{\text{sMAPE}(A,i) - \text{sMAPE}(B,i)}{\text{sMAPE}(A,i)}.$$

Negative Δ indicates Model A achieves lower error (better performance) than Model B.

The solid red curve is a LOWESS (locally weighted scatterplot smoothing) fit of $\Delta(A \to B)$ as a function of Ω, using a smoothing fraction of 0.4. The shaded red region is an empirical 95% confidence band for that trend, constructed via a nonparametric bootstrap: we resample datasets with replacement 300 times, recompute the LOWESS fit for each bootstrap resample, and evaluate each fitted curve on a common (Ω) grid. For each (Ω) value on that grid, we take the 2.5th and 97.5th percentiles across the bootstrap fits; these percentiles define the lower and upper edges of the shaded band. Thus, the solid red line is a visualizer that shows the smoothed observed relationship between (Ω) and relative error gain, and the translucent band shows the bootstrap variability of that relationship across datasets.

Fig. 6 reveals a striking pattern: **zero-shot models show systematically increasing advantages as Ω increases**, with performance gains reaching 20–60% over statistical and deep learning baselines in high-Ω regimes ($\Omega > 0.5$). This relationship is moderately consistent across all three comparisons (vs. statistical, vs. deep learning, vs. pretrained), with Spearman correlations ranging from -0.234 for statistical to -0.556 for deep learning.

Critically, this advantage is *specific to zero-shot models*. Fig. 7 shows that pretrained models exhibit no consistent Ω-dependent pattern when compared to statistical or deep learning baselines. Their relative performance displays minimal trend across the Ω spectrum, suggesting that the process of "re-pretraining" TSFMs somehow degraded their ability to exploit the highest-Ω ranges. This suggests that the pretraining corpus—not just model architecture or size—fundamentally shapes how models respond to spectral structure, a promising future area of investigation.

The Low-Ω Challenge. Critically, as Ω decreases below 0.2, performance differences between models narrow substantially. At $\Omega < 0.2$, there is minimal differentiation between between model types; all struggle similarly. This reveals an important research gap: current models, whether lightweight or sophisticated, have made limited progress on genuinely difficult (low-Ω) problems. The field has optimized performance on predictable data without developing architectures that push advantages where forecasting is hardest.

Practical Implications for Model Selection

These results provide clear, actionable guidance for practitioners:

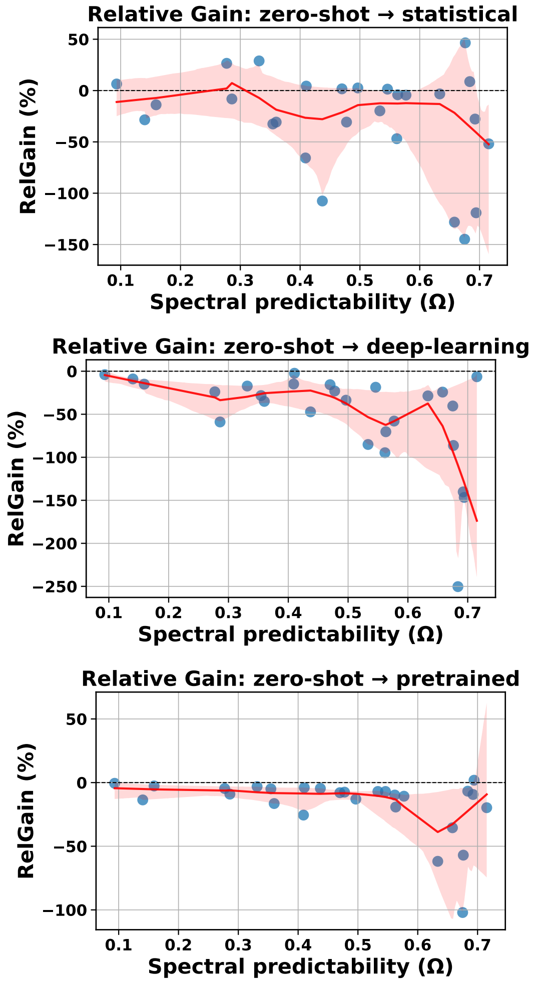

Figure 6: **Zero-shot models uniquely exploit high-Ω regimes.** Negative values indicate zero-shot models achieve lower error. The average performance advantage improves with Ω.

High-Ω Datasets ($\Omega > 0.5$)**:** Zero-shot models are the optimal choice. A practitioner who computes Ω and finds their dataset falls in this regime can narrow their search space to zero-shot candidates (e.g., Moirai or Chronos) rather than validating a multitude of diverse models. This reduces the computational costs of validation substantially while maintaining or improving expected performance.

Low-Ω Datasets ($\Omega < 0.4$)**:** Zero-shot models provide no clear advantage. Simpler statistical or deep learning models (like ARIMA or DLinear, respectively) offer comparable accuracy with substantially lower computational cost (both training and inference). For resource-constrained deployments or when rapid iteration is needed, this tradeoff strongly favors lightweight alternatives.

Mid-Ω Datasets ($0.4 \leq \Omega \leq 0.5$)**:** The choice is less clear-cut. Practitioners should consider additional factors: computational budget, inference latency requirements, and whether the dataset exhibits characteristics (missingness, regime shifts) that Ω may not fully capture.

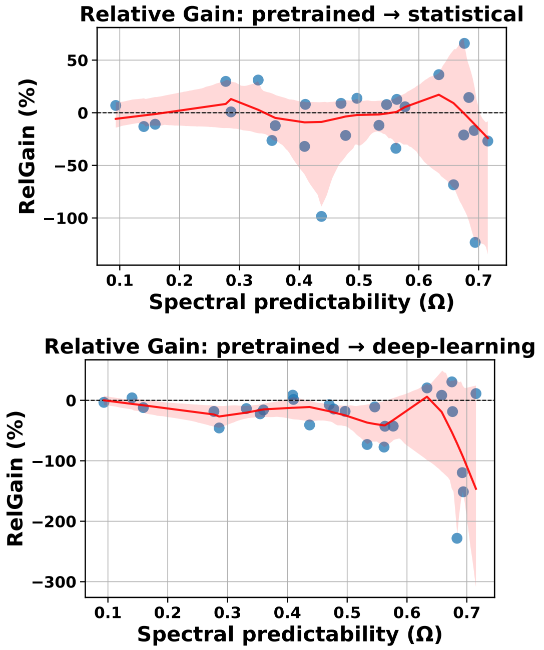

Figure 7: **Pretrained models show no systematic Ω-dependence.** Unlike zero-shot models, pretrained models do not exhibit predictable advantages based on spectral predictability.

When to Trust Ω: Reliability Conditions

While Ω provides useful guidance across diverse datasets, its reliability depends on specific data characteristics. We identify three conditions that limit Ω's predictive power:

(1) Non-stationary processes. Ω assumes spectral properties remain stable over time. For series with regime shifts or structural breaks, or those with high LLE, a single aggregate Ω may not reflect local forecasting difficulty.

(2) Insufficient data length. FFT-based spectral estimates become unstable for short series (we recommend $T > 1000$ timesteps).

(3) Exogenous-driven dynamics. Ω characterizes intrinsic temporal structure. For processes dominated by external shocks (e.g., traffic accidents, policy interventions), spectral properties may not capture true predictability. The weaker patterns in PEMS and Fitbit (Fig. 1) likely reflect such factors.

Failure case example. In Fig. 6, an outlier dataset point exhibits $\Omega = 0.66$ but statistical models outperform zero-shot by nearly 50%, counter to expectations. Further analysis is needed to understand dynamics not fully captured by Ω.

Decision heuristic. Before applying Ω-based selection, practitioners should verify: (i) series length is sufficient for a rigorous computation of Ω, (ii) visual inspection or statistical tests suggest stationarity, and (iii) domain knowledge indicates intrinsic temporal patterns dominate exogenous shocks. When these conditions hold, Ω provides reliable guidance; otherwise, complement with domain-specific heuristics.

Discussion

Why Not Just Use Validation Error?

A natural question arises: why not simply train all candidate models and pick the one with lowest validation error? We argue spectral predictability offers three key advantages:

(1) Computational Efficiency. Computing Ω takes seconds; training and validating multiple models can take hours to days. For a practitioner with many candidate models, our approach reduces the search space *before any training begins*, saving substantial compute and engineering time. This represents meaningful cost savings, especially when iterating across multiple datasets or deployment scenarios.

(2) Interpretability and Generalization. Validation error reports *which* model performed best on a specific dataset, but not *why*. Ω provides interpretable signal properties that generalize across datasets. If a practitioner encounters a new high-Ω dataset, they can leverage prior knowledge that zero-shot models excel in this regime without re-running exhaustive validation. This builds intuition and enables faster decision-making.

(3) Highlighting Research Gaps. Validation error alone doesn't reveal systematic patterns in model behavior. By stratifying performance along Ω, we expose that models primarily differentiate on easy problems (high-Ω) while converging on hard ones (low-Ω). This insight motivates targeted research into improving performance where it matters most.

In practice, we envision Ω as a *first-pass filter* that complements, rather than replaces, validation. The decision process: compute Ω, narrow the model space based on regime, then validate the reduced candidate set. This hybrid approach balances efficiency with empirical rigor.

Limitations

Multivariate or Nonstationary Data. In this work we define Ω for univariate series, and for multivariate datasets we summarize by averaging per-channel Ω. This is a pragmatic choice but does not capture joint temporal structure (e.g., cross-channel coherence). A more principled extension would compute joint spectral predictability using cross-spectral densities or coherence-based summaries, or by applying Ω to low-dimensional projections (e.g., principal components) that capture shared dynamics. Similarly, for non-stationary series with regime shifts, a single global spectrum can be misleading; a natural extension is a *local* or *time-varying* Ω computed over sliding windows or with time-frequency methods, enabling practitioners to detect predictability changes over time and adapt model choice accordingly.

Other Correlational Effects. Though our exploration makes a case for Ω as a predictor of model performance, our preliminary results suggest possible effects on accuracy from both the pretraining corpus of the TSFM selected and the LLE of the data. More work should be done to integrate these indicators into a more cohesive index.

Model coverage. GIFT-Eval provides broad coverage but may not represent all architecture types. Notably absent are neural ODEs and certain probabilistic models that might exhibit different Ω relationships.

Aggregation effects. Computing Ω on full datasets when train-test splits are unavailable creates potential for subtle aggregation bias. Our controlled experiments (where Ω is computed per-series on test data only) replicate key patterns, suggesting findings are robust, but future work should examine sensitivity to this choice.

Conclusion and Future Work

Our key findings provide actionable guidance: for high-Ω datasets ($\Omega > 0.5$), zero-shot models generally outperform alternatives; for low-Ω datasets ($\Omega < 0.4$), zero-shot advantages disappear and lightweight models offer comparable accuracy at lower cost. We identify when Ω is reliable (stationary processes, sufficient length) and when practitioners should complement it with domain knowledge (regime shifts, exogenous shocks).

Beyond practical model selection, our analysis reveals that current models differentiate primarily on easy (high-Ω) problems, with performance gaps narrowing substantially as Ω decreases. This highlights a critical research need: developing architectures that excel specifically on genuinely difficult forecasting challenges. We propose specific directions including attention mechanisms for sparse pattern detection, mixture-of-experts for adaptive capacity allocation, and explicit noise modeling.

Future work can extend this framework along several directions. First, Ω could play a central role in *agentic time-series AI*: autonomous systems that monitor their data streams, estimate Ω in real time, and self-configure by choosing lightweight or foundation-scale models as conditions change. Second, integrating Ω as a standardized metadata field in future benchmarks would promote reproducibility and allow systematic comparisons of model reliability across predictability regimes. Third, our findings highlight the need to advance modeling in the *low-Ω regime*, where current architectures converge in performance. Finally, generalizing (Ω) to *multivariate and multimodal* time series would enable a notion of joint predictability that captures more complex dependencies.

Together, these directions outline a vision of predictability-aware AI systems that use spectral structure not only to guide model selection but also to drive self-adaptive, resource-efficient, and interpretable time-series intelligence.

Acknowledgements

The research reported in this paper was sponsored in part by the National Science Foundation (awards # CNS-2211301 and CNS-232595), the National Institutes of Health (award # 1P41EB028242), and the DEVCOM Army Research Laboratory (award # W911NF1720196). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the funding agencies. Mani Srivastava was also partially supported by the Mukund Padmanabhan Term Chair at UCLA.

References

Aksu, T.; Woo, G.; Liu, J.; Liu, X.; Liu, C.; Savarese, S.; Xiong, C.; and Sahoo, D. 2024. GIFT-Eval: A Benchmark For General Time Series Forecasting Model Evaluation. arXiv:2410.10393.

Ansari, A. F.; Stella, L.; Turkmen, C.; Zhang, X.; Mercado, P.; Shen, H.; Shchur, O.; Rangapuram, S. S.; Arango, S. P.; Kapoor, S.; Zschiegner, J.; Maddix, D. C.; Wang, H.; Mahoney, M. W.; Torkkola, K.; Wilson, A. G.; Bohlke-Schneider, M.; and Wang, Y. 2024. Chronos: Learning the Language of Time Series. arXiv:2403.07815.

Elsayed, S.; Thyssens, D.; Rashed, A.; Jomaa, H. S.; and Schmidt-Thieme, L. 2021. Do We Really Need Deep Learning Models for Time Series Forecasting? arXiv:2101.02118.

Furberg, R.; Brinton, J.; Keating, M.; and Ortiz, A. 2016. Crowd-sourced Fitbit Datasets (03/12/2016–05/12/2016). Dataset.

Goswami, M.; Szafer, K.; Choudhry, A.; Cai, Y.; Li, S.; and Dubrawski, A. 2024. MOMENT: A Family of Open Time-series Foundation Models. arXiv:2402.03885.

Gruver, N.; Finzi, M.; Qiu, S.; and Wilson, A. G. 2024. Large Language Models Are Zero-Shot Time Series Forecasters. arXiv:2310.07820.

Guntu, R. K.; Yeditha, P. K.; Rathinasamy, M.; Perc, M.; Marwan, N.; Kurths, J.; and Agarwal, A. 2020. Wavelet entropy-based evaluation of intrinsic predictability of time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(3).

Jin, M.; Wang, S.; Ma, L.; Chu, Z.; Zhang, J. Y.; Shi, X.; Chen, P.-Y.; Liang, Y.; Li, Y.-F.; Pan, S.; and Wen, Q. 2024. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. arXiv:2310.01728.

Li, Z.; Qiu, X.; Chen, P.; Wang, Y.; Cheng, H.; Shu, Y.; Hu, J.; Guo, C.; Zhou, A.; Jensen, C. S.; et al. 2025. TSFM-Bench: A comprehensive and unified benchmark of foundation models for time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 5595–5606.

Liang, Y.; Wen, H.; Nie, Y.; Jiang, Y.; Jin, M.; Song, D.; Pan, S.; and Wen, Q. 2024. Foundation Models for Time Series Analysis: A Tutorial and Survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, 6555–6565. ACM.

Maji, D.; Shenoy, P.; and Sitaraman, R. K. 2022. CarbonCast: multi-day forecasting of grid carbon intensity. In *Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, 198–207.

Miller, J. A.; Aldosari, M.; Saeed, F.; Barna, N. H.; Rana, S.; Arpinar, I. B.; and Liu, N. 2024. A Survey of Deep Learning and Foundation Models for Time Series Forecasting. arXiv:2401.13912.

Pincus, S. 1991. Approximate entropy as a measure of system complexity.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners.

Shi, J.; Ma, Q.; Ma, H.; and Li, L. 2024. Scaling Law for Time Series Forecasting. arXiv:2405.15124.

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. arXiv:2409.16040.

Tan, M.; Merrill, M. A.; Gupta, V.; Althoff, T.; and Hartvigsen, T. 2024. Are Language Models Actually Useful for Time Series Forecasting? arXiv:2406.16964.

Tang, H.; Zhang, C.; Jin, M.; Yu, Q.; Wang, Z.; Jin, X.; Zhang, Y.; and Du, M. 2024. Time Series Forecasting with LLMs: Understanding and Enhancing Model Capabilities. arXiv:2402.10835.

Wang, R.; Klee, S.; and Roos, A. 2025. Time Series Forecastability Measures. arXiv:2507.13556.

Wang, Y.; Wu, H.; Dong, J.; Liu, Y.; Long, M.; and Wang, J. 2024. Deep Time Series Models: A Comprehensive Survey and Benchmark.

Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; and Long, M. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. arXiv:2210.02186.

Ye, J.; Zhang, W.; Yi, K.; Yu, Y.; Li, Z.; Li, J.; and Tsung, F. 2024. A Survey of Time Series Foundation Models: Generalizing Time Series Representation with Large Language Model. arXiv:2405.02358.

Zeng, A.; Chen, M.; Zhang, L.; and Xu, Q. 2022. Are Transformers Effective for Time Series Forecasting? arXiv:2205.13504.

Appendix

Model Categorization by Type

For completeness, we list below which models fall under each model type used in our analyses. Note that similar models in the same family may have been applied differently (e.g., used with pretraining versus in zero-shot). The models labeled as "agentic" have an emphasis on dynamic ensemble models.

GIFT-Eval Zero-shot models: TimesFM-2.5, FlowState-9.1M, Kairos_10m, Kairos_23m, Kairos_50m, sundial_base_128m, YingLong_110m, YingLong_300m, YingLong_50m, YingLong_6m, granite-flowstate-r1, TiRex, DeOSAlphaTimeGPTPredictor-2025, Toto_Open_Base_1.0, TabPFN-TS, Moirai_base, Moirai_large, Moirai_small, and VisionTS. **Pretrained models:** Moirai2, Chronos_bolt_base, Chronos_bolt_small, Chronos_large, Chronos_base, Chronos_small, TimesFM, timesfm_2_0_500m, TTM-R1-Pretrained, TTM-R2-Pretrained, and Lag-Llama. **Deep learning models:** xLSTM-Mixer, PatchTST, TFT, N-BEATS, DLinear, TIDE, DeepAR, Crossformer, and iTransformer. **Statistical baselines:** Auto_Arima, Seasonal_Naive, Auto_Theta, Auto_ETS, and Naive. **Fine-tuned models:** TSOrchestra-test, TTM-R2-Finetuned, and TEMPO_ENSEMBLE. **Agentic models:** Kairos-1.0, Nexus-1.0, TSOrchestra, and TimeCopilot.

Dataset Protocol

Cross-series generalization. Models are trained on heterogeneous series within each domain, spanning a wide range of spectral entropies, and evaluated on held-out series covering similar ranges. This enables assessment of how well models generalize to unseen series with varying Ω values.

For real-world datasets, we select training series to represent the full range of spectral entropies available in each domain. CarbonCast and PEMS are binned into low, medium, and high tertiles, with 6 series randomly chosen from each bin (18 total). For Fitbit, which has fewer users, we select 5 series at each quintile. Test data come from held-out series at the 5th, 50th, and 95th percentiles of spectral entropy. Note that this explains why the entropy range within each figure's subplots may vary.

Synthetic training data sweep 8 spectral entropy levels from 0.25 to 0.85 in increments of 0.10; test data cover 0.2 to 0.8 with the same intervals.

PEMS and CarbonCast originate as multivariate datasets. PEMS is aggregated to hourly resolution (following Wang et al. (2024)), and all covariates are split into independent univariate series to enable spectral entropy computation and avoid covariate learning effects. These series are treated as same-domain data with distribution shifts (*e.g.*, across regions in PEMS or energy sources in CarbonCast). For Fitbit, user series with missing (zero) values in the test window are excluded.

GIFT-Eval Conversion For the original datasets present in GIFT-Eval's analysis, we calculated metrics such as Ω and LLE per each covariate, and averaged them together for a multivariate dataset. Due to runtime constraints and some very large datasets, we truncated the incoming data at 4096 steps per data stream. To calculate LLE, we reconstructed the state space with an embedding dimension $m = 4$ and delay $\tau = 10$ steps. We then estimated the LLE using standard Rosenstein-style local divergence tracking. These hyperparameters were chosen to ensure sufficient samples for a robust estimate without taking prohibitively long, and further investigation can be done here.

Training Protocol

Single H100 GPU, three seeds per model. LR $= 0.01$, batch $= 16$, up to 10 epochs with early stopping (patience 3). Inputs normalized per context window; stride 1; boundary regime prevents window crossing across different series.

Spectral Predictability Details

We apply a Hann taper window for a balanced compromise between main-lobe width and side-lobe suppression to reduce spectral leakage. We choose this because we analyze diverse time series whose periodicities are not known in advance. For controlled experiments, we compute Ω from each test series (using only the input sequence, not forecast targets) to characterize the spectral properties each model must forecast. This avoids leakage while providing a per-series difficulty measure.

Table 3: Correlation statistics between Ω (spectral predictability) and sMAPE. 95% CIs computed via Fisher z-transform. Narrow intervals reflect low variability across seeds.

Statistic	**CarbonCast**	**Fitbit**	**PEMS**	**Synthetic**
n	33	18	18	42
Pearson r	-0.68	-0.38	-0.75	-0.72
95% CI (low)	-0.83	-0.72	-0.90	-0.84
95% CI (high)	-0.43	-0.11	-0.43	-0.53
Spearman ρ	-0.74	-0.38	-0.71	-0.68

Table 4: Aggregate relationship between Ω and MSE. Negative slopes indicate that forecasting error decreases as predictability increases, supporting Ω as a proxy for difficulty.

Model	Median slope
Aggregate	-1.08
DLinear	-0.97
GPT2	-1.08
Language Pretrained	-1.17
Random Init	-1.09

Statistical Results of Controlled Experiments

In addition to our controlled experiment conclusions identified earlier, we report the corresponding tabular results in MSE, though similar trends hold for sMAPE. Aggregate statistics are computed as the mean of the domain-specific values for clarity and comparability.

(i) Forecasting Error Decreases as Predictability Increases. Table 4 reports the aggregate relationship between MSE and spectral predictability (Ω) across models. Across all settings, the median slopes for MSE are also negative, suggesting that error systematically declines as predictability rises. This trend holds for lightweight models such as DLinear as well as for larger pretrained backbones, reinforcing Ω as a proxy for forecasting difficulty. Notably, DLinear shows the smallest magnitude of slope, reflecting its strong performance in predictable regimes, while language-pretrained models and random initialization exhibit slightly steeper slopes, indicating greater sensitivity to Ω.

(ii) DLinear Dominates in Predictable Regimes. Table 5 presents the Theil–Sen slope of the relative Error Increase Δ (%) versus Ω, comparing Language Pretrained models against DLinear. Positive slopes indicate that the performance gap in favor of DLinear widens as predictability increases, suggesting a strong inductive bias for structured series. The table also reports Spearman ρ and Pearson r correlations to assess robustness. Results show consistent positive slopes in CarbonCast, PEMS, and Synthetic, with strong correlations ($r > 0.85$), confirming that DLinear outperforms pretrained models in high-Ω regimes. The aggregate trend is likewise positive, while Fitbit deviates with a negative slope and weak correlations, highlighting the unique irregularities of wearable data. Overall, these results reinforce that DLinear is most effective when time series exhibit strong seasonality or repetition, while pretrained models gain relevance as predictability falls.

Table 5: Theil–Sen slope of Error Increase Δ (%) *vs.* Ω for Language Pretrained relative to DLinear. Positive slopes indicate that performance gap in favor of DLinear widens as predictability increases.

Domain	Slope	Spearman ρ	Pearson r
Aggregate	87.97	0.57	0.58
CarbonCast	225.41	1.00	0.95
Fitbit	-47.23	-0.50	-0.47
PEMS	72.73	1.00	0.96
Synthetic	100.97	0.79	0.87

Table 6: Theil–Sen slope of Error Increase Δ (%) versus Ω for Language Pretrained relative to GPT-2, showing how model size affects performance across predictability levels.

Domain	Slope	Spearman ρ	Pearson r
Aggregate	-2.08	0.32	0.10
CarbonCast	69.99	1.00	0.94
Fitbit	-88.34	-0.50	-0.87
PEMS	7.27	0.50	0.19
Synthetic	2.76	0.29	0.15

(iii) Model Size Has Limited Effect Across Predictability Levels. Table 6 is structured similarly to Table 5, but compares Language Pretrained models to GPT-2 (a smaller backbone). It evaluates how model size influences relative performance when conditioned on spectral predictability Ω. The aggregate slope is near zero (-2.08), indicating little systematic advantage for either model overall. CarbonCast shows a strongly positive slope (69.99) with high correlations ($r = 0.94$), suggesting that larger pretrained models perform better on high-Ω energy data. In contrast, Fitbit again shows a strongly negative slope (-88.34), reflecting its irregular missingness and confirming it as an exception. PEMS and Synthetic display small, weakly positive slopes with low correlations, reinforcing that scaling effects are inconsistent across domains. Overall, these results suggest that while model size may provide gains in certain settings, Ω remains the dominant axis for explaining relative performance, and scaling alone does not guarantee improvements across all domains.

(iv) Pretraining Provides Limited Gains Beyond the Embedding Head. Table 7 follows the structure of Table 5, but compares Language Pretrained models to Random Init, isolating the effect of pretraining. It evaluates how pretraining influences relative performance across levels of spectral predictability Ω. The aggregate slope is small (15.8) with negligible correlations, suggesting limited systematic benefit of pretraining overall. CarbonCast stands out with a strong positive slope (127.64, $r = 0.94$), implying that pretraining can help in highly structured energy data. By contrast, Fitbit again shows a large negative slope (-68.15) with strong negative correlations, reinforcing it as an outlier due to irregular missingness. PEMS and Synthetic display slopes close to zero with weak correlations, further supporting the view that pretraining offers little advantage once model capacity is held constant. Overall, these results suggest that much of the forecasting ability stems from the large embedding head itself, with pretraining only adding value in select domains.

Table 7: Theil–Sen slope of Error Increase Δ (%) versus Ω for Language Pretrained relative to Random Init, isolating the effect of pretraining.

Domain	Slope	Spearman ρ	Pearson r
Aggregate	15.8	0.13	0.01
CarbonCast	127.64	1.00	0.94
Fitbit	-68.15	-1.00	-1.00
PEMS	3.96	0.50	0.14
Synthetic	-0.28	0.00	-0.03

Instructive Exceptions

The Fitbit and PEMS domains show weaker alignment between error and Ω compared to other datasets. We attribute this to the sparsity of the two datasets. In wearable data, users often remove devices for extended periods (*e.g.*, during sleep), creating irregular gaps. In traffic data, there are stretches of time, especially at night, where no vehicles arrive at a given intersection sensor. Because Ω is a purely spectral measure, it does not fully capture these missingness patterns, which may dominate the forecasting difficulty for some of the tasks chosen. A more thorough analysis, such as filtering for active periods, applying imputation strategies, or testing alternative error metrics, is an interesting direction for future work.