Evaluation of AI-based Forecasting Models for Electricity Demand at Household Level: Focus on Generative AI Models

Erwan Chibout, Simon Camal, Georges Kariniotakis

EVALUATION OF AI-BASED FORECASTING MODELS FOR ELECTRICITY DEMAND AT HOUSEHOLD LEVEL: FOCUS ON GENERATIVE AI MODELS

***Erwan Chibout*[1]*, *Simon Camal*[1], *Georges Kariniotakis*[1]**

[1]Mines Paris – PSL, Centre PERSEE, Sophia-Antipolis, France
*erwan.chibout@minesparis.psl.eu

Keywords: TIME SERIES FOUNDATION MODELS, RESIDENTIAL ENERGY FORECASTING, ZERO-SHOT LEARNING, BENCHMARKING, COMPUTATIONAL EFFICIENCY

Abstract

Short-term forecasting of electricity consumption, from a few minutes to day ahead, is critical for distribution system operators (DSOs) to manage the electricity grid, balance supply and demand, and integrate renewable energy sources. For a long time, forecasting approaches used to rely on statistical methods (ARIMA, exponential smoothing) or classical "lightweight" machine learning models (Gradient Boosting, Random Forest). These models require extensive feature engineering and show limited accuracy and robustness. The recent emergence of Time Series Foundation Models (TSFMs), mostly based on transformer architectures and pre-trained on millions of time series are capable of "zero-shot" forecasting, which means without specific training, offers a promising alternative for forecasting on unseen data . However, their practical applicability to specific tasks such as residential consumption forecasting is still poorly evaluated, particularly regarding forecasting accuracy and computational efficiency. This work presents a comprehensive benchmark comparing 12 models, from naive baselines to state-of-the-art TSFMs, on residential electricity consumption data at 15-minute resolution. Results show that fine-tuned TSFMs achieve the lowest NRMSE (0.58 ± 0.14), outperforming classical ML approaches (XGBoost: NRMSE 0.65), while zero-shot variants remain competitive without any task-specific training. The high MAPE values (≈45–75 %) observed across the models is mainly due to the nature of the forecasting at household scale.

1 Introduction

The accelerating deployment of distributed energy resources (rooftop solar panels, electric vehicles, heat pumps) and the growing electrification of uses are changing distribution network operating conditions. The load profiles at residential level are becoming more and more volatile and uncertain, making accurate short-term demand forecasting a critical operational requirement for DSOs and local aggregators constraint for renewable integration and demand response coordination.[1]

Classical approaches to residential load forecasting have relied on statistical time series models such as ARIMA or exponential smoothing, or on supervised machine learning algorithms such as gradient boosting (XGBoost, LightGBM) and random forests have shown some limits. Deep learning architectures (LSTM, GRU, Transformers) have demonstrated improvements but still demand large amounts of historical data coupled with an important model tuning.[2]

These issues are adressed by Time Series Foundation Models (TSFMs): large-scale pre-trained models capable forecasting on unseen time series without task-specific training. Models such as Chronos (Amazon)[3], TimesFM (Google)[4], and LLMTime (language model-based)[5] have shown competitive performance on standard benchmarks. This work intends to evaluate their on residential electricity data at high temporal resolution (15-minute granularity).

1.1. Related Work

Several studies have benchmarked machine learning models for short-term load forecasting at the aggregate level (distribution feeders or substations) where the signal is smoother and more predictable. At the household level, however, consumption shows higher stochasticity due to occupant behaviour and appliance usage patterns, making accurate forecasting considerably more challenging [6]. Recent TSFM benchmarks such as the GIFT-Eval suite or the ETDataset evaluations have documented strong zero-shot performance across diverse domains, but residential electricity data is systematically underrepresented in these assessments.

1.2. Research Objectives

This paper addresses the following question: *Do Time Series Foundation Models offer a practical advantage over classical approaches for day-ahead residential electricity demand forecasting, and at what computational cost?*

This question is motivated by three observations. First, individual household consumption is intrinsically stochastic, dominated by appliance switching events and occupant behaviour rather than smooth temporal patterns [7], which makes it structurally different from the aggregate or commercial loads on which deep learning models are already able to provide accurate and robust forecasts. Second, TSFMs are increasingly used as zero-shot or few-shot solutions, yet their behaviour under fine-tuning on small residential datasets has not been systematically characterised. Third, DSOs deploying forecasting systems at scale — across hundreds or thousands of smart meters — require models whose inference time remains tractable, a dimension largely absent from academic benchmarks.

To address these gaps, we provide a comprehensive bench-

mark comparing 12 forecasting approaches across three categories: classical statistical baselines (Persistence, ARIMA), trained machine learning models (XGBoost, TimeGAN), and TSFMs evaluated both in zero-shot and fine-tuned regimes (Chronos-Mini, Chronos-Small, Chronos-Bolt, TimesFM, LLMTime-Mistral, Autoformer[8]). All models are evaluated on a standardised residential dataset under identical conditions, with explicit and systematic reporting of both forecasting accuracy (NRMSE, MAPE) and inference time per forecast horizon.

We aim to provide guidance on model selection for operational deployment, and to establish realistic performance baselines for individual-meter residential forecasting.

2 Methodology

2.1. Benchmark Design

The benchmark spans three categories of models:

2.1.1. Naive and statistical baselines:

- **Persistence**: the forecast for day D equals the observed consumption of day $D-1$; used as a lower-bound reference.
- **ARIMA**: autoregressive integrated moving average with automatic order selection, trained via rolling-window cross-validation.

2.1.2. Classical machine learning:

- **XGBoost**: gradient boosting with 48 engineered features (temporal lags, rolling statistics, calendar indicators). Trained on historical consumption data specific to each household.
- **GAN-based model (TimeGAN)**: generative adversarial network architecture for time series generation, assessed as a data-augmentation-aware forecaster.

2.1.3. Time Series Foundation Models:

- **Chronos-2 variants** (Amazon): Mini (20M), Small (46M), and Bolt (50M, speed-optimised), evaluated in both zero-shot and fine-tuned configurations.
- **TimesFM** (Google Research, 200M parameters): a decoder-only foundation model for time series, zero-shot only.
- **Autoformer**: Transformer-based architecture with auto-correlation mechanism, evaluated fine-tuned.
- **LLMTime with Mistral-7B**: language model repurposed for time series forecasting by encoding numerical values as text tokens, zero-shot configuration.

2.2. Dataset

The benchmark uses residential electricity consumption data from the Open Power System Data (OPSD) platform, covering 11 German households from the Konstanz region recorded at 15-minute temporal resolution ($\Delta t = 15$ min). Training data spans 2015–2018 (≈140,000 timesteps per household); evaluation is performed on the held-out year 2019 (35,040 timesteps). Missing values, present in less than 1.2% of all records, are imputed by linear interpolation prior to model training and evaluation.

The dataset exhibits characteristics representative of residential forecasting challenges. Intra-day variability is pronounced, with consumption ranging from near-zero baseload at nighttime ($C \approx 0.05$–0.15 kWh/15 min) to morning and evening peaks ($C \approx 0.4$–1.2 kWh/15 min). Inter-household heterogeneity is substantial: mean daily consumption varies from 4.1 kWh to 18.7 kWh across households (coefficient of variation ≈ 0.62), reflecting differences in household size and occupancy schedules. All 11 households are also showing strong weekly seasonality, with weekend profiles systematically deviating from weekday patterns, which is expected in this context. These properties make the dataset well-suited for benchmarking models across a realistic diversity of residential consumption behaviours.

2.3. Problem Formulation

Let $\mathcal{H} = \{1, \ldots, N\}$ with $N = 11$ denote the set of households, and let $C_{i,t} \in \mathbb{R}_+$ denote the electricity consumption of household $i \in \mathcal{H}$ at time t, measured in kWh per 15-minute interval. Residential consumption is decomposed as:

$$C_{i,t} = \mu_i(t) + \varepsilon_{i,t} \tag{1}$$

where $\mu_i(t)$ captures calendar seasonality (time of day, day of week, month) and $\varepsilon_{i,t}$ is a zero-mean stochastic residual encoding appliance-level randomness and occupant behaviour.

The day-ahead forecasting task consists in estimating the full consumption profile of day D given all information available at forecast time $t^* = D - 12$ h (12:00 on day $D-1$):

$$\hat{C}_{i,t^*+h} = f_\theta\left(C^{(L)}_{i,t^*},\, X_{t^*}\right), \quad h = 1, \ldots, H \tag{2}$$

where $C^{(L)}_{i,t^*} = \{C_{i,t^*-L+1}, \ldots, C_{i,t^*}\}$ is the lookback window of $L = 672$ timesteps (7 days), $H = 96$ is the forecast horizon (24 hours at 15-minute resolution), and X_{t^*} denotes optional exogenous features (calendar indicators, rolling mean and variance). Effective forecasting horizons range from 12 h to 36 h ahead. For fine-tuned and classical ML models, θ is estimated by minimising a mean squared error loss on $\mathcal{T}_{\text{train}}$ (2015–2018), independently per household. For zero-shot foundation models, θ is fixed from pre-training; no household-specific optimisation is performed.

2.4. Evaluation Protocol

All models are evaluated in a strict day-ahead rolling forecast setting: for each day D of the evaluation year 2019, a forecast is produced using only information available at $t^* = D - 12$ h, This protocol enables to avoid data leakage: test-period observations are never accessible to any model, and for fine-tuned models, hyperparameter selection is performed via cross-validation on the 2015–2017 subset, with 2018 serving as a validation year.

Evaluation metrics: Two complementary metrics are reported. **NRMSE** (Normalized Root Mean Square Error) is computed by dividing RMSE by the standard deviation σ of observed consumption over the evaluation period: NRMSE = $\frac{\text{RMSE}}{\sigma}$, which enables comparison across households with different absolute consumption levels and penalises large deviations

more heavily, making it particularly sensitive to peak forecasting errors. **MAPE** (Mean Absolute Percentage Error) provides a scale-independent relative error that gives equal weight to all timesteps regardless of consumption magnitude; however, near-zero nighttime values can inflate it significantly and results should therefore be interpreted alongside NRMSE. Inference time per 24-hour forecast is additionally reported as an indicator for operational deployability.

Results are reported as mean ± standard deviation across the 11 households to capture inter-household variability.

3 Results and Analysis

3.1. Forecasting Accuracy

Table 1 summarises NRMSE performance across all models and configurations.

Model	Config	NRMSE	MAPE (%)
Persistence	–	0.98 ± 0.22	74.3 ± 20.1
ARIMA	–	0.84 ± 0.19	62.8 ± 17.4
XGBoost	Trained	0.65 ± 0.16	49.6 ± 14.1
TimeGAN	Trained	0.77 ± 0.18	57.3 ± 15.6
Chronos-Mini (20M)	Zero-shot	0.73 ± 0.17	55.2 ± 14.3
Chronos-Small (46M)	Zero-shot	0.67 ± 0.16	50.8 ± 13.5
Chronos-Bolt (50M)	Zero-shot	0.66 ± 0.15	50.1 ± 13.2
TimesFM (200M)	Zero-shot	0.69 ± 0.16	52.4 ± 13.8
LLMTime-Mistral	Zero-shot	0.74 ± 0.17	56.7 ± 14.5
Chronos-Small (46M)	Fine-tuned	**0.58 ± 0.14**	**44.9 ± 12.3**
Autoformer	Fine-tuned	0.62 ± 0.15	47.8 ± 13.1

Table 1: Day-ahead forecasting performance (NRMSE and MAPE, mean ± std across 11 households). NRMSE normalised by the std of observed consumption; MAPE inflated by near-zero nighttime values at 15-min resolution.

Foundation models demonstrate competitive performance with significantly reduced preprocessing requirements compared to traditional approaches. Chronos-Small (fine-tuned) achieves the lowest NRMSE (0.58 ± 0.14) and MAPE (44.9 ± 12.3%), outperforming XGBoost (NRMSE: 0.65, MAPE: 49.6%) and traditional statistical methods (ARIMA: NRMSE 0.84, MAPE 62.8%). Zero-shot foundation models (TimesFM, Chronos-Bolt) maintain strong performance (NRMSE ≈ 0.66–0.73, MAPE ≈ 50–55%) without any task-specific training, confirming the generalisation capability of large pre-trained architectures.

The high error levels across all models are consistent the expectations individual household forecasting : at the single-meter level, consumption is dominated by stochastic appliance switching and occupant behaviour. These values stand in stark contrast to aggregate-level forecasting where MAPE below 5% is routinely achieved.

The relatively high standard deviations across all models (e.g., ±0.14 for the best configuration) reflect the strong inter-household heterogeneity: households with irregular occupancy or atypical appliance usage are systematically harder to forecast for all model classes.[9] MAPE values are consistently higher than expected from NRMSE alone, given the nighttime inflation effect: periods with $C_{i,t} \approx 0$ generate relative errors that imply an increase of the household-level mean, particularly for households with more pronounced nocturnal baseload suppression in the dataset.

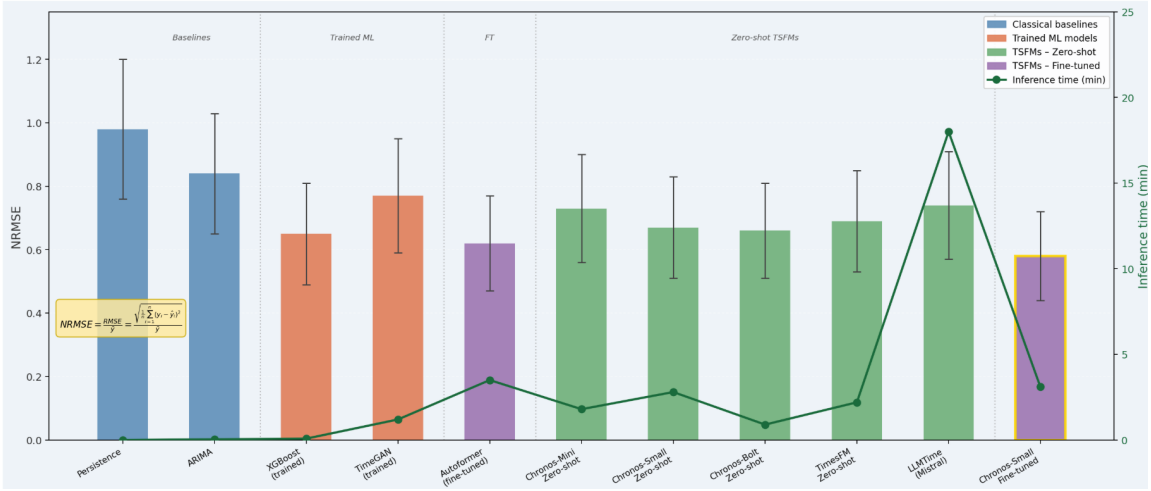

Figure 1: Comparison of NRMSE and inference time (CPU) across 11 forecasting approaches for day-ahead residential electricity demand forecasting at 15-minute resolution. Error bars represent standard deviation across 11 households. NRMSE normalised by the standard deviation of observed consumption.

3.2. Computational Efficiency

Computational efficiency varies dramatically across models. Lightweight TSFM variants such as Chronos-Bolt (zero-shot) achieve near-competitive accuracy in under 1 minute for the full household group on standard CPU hardware, making them the most operationally attractive option when GPU infrastructure is unavailable. Chronos-Small and TimesFM require 2 to 3 minutes per forecast cycle, remaining viable for daily batch-mode deployment at moderate scale.

LLMTime-Mistral, however, requires approximately 18 minutes per forecast on CPU — an order of magnitude higher than specialised TSFMs, with no corresponding accuracy gain. This effectively disqualifies LLM-based approaches for operational deployment unless dedicated GPU resources are available.

More broadly, while zero-shot TSFMs eliminate training overhead, their inference cost scales linearly with the number of meters. Deployment across thousands of smart meters, a realistic DSO scenario, would require 10 to 50 hours of cumulative CPU inference per day for the larger models, making GPU acceleration no longer optional but necessary. On a modern GPU (NVIDIA A100), inference times are expected to decrease by one to two orders of magnitude restoring the operational viability of heavier architectures. Fine-tuning costs, not reported here, represent an additional one-time overhead that should be factored into total cost of ownership assessments. It confirms that these models are suited for operational deployment [10]

Classical baselines (Persistence, ARIMA, XGBoost) remain unmatched in terms of inference speed, completing forecasts in seconds regardless of scale, which preserves their relevance in resource-constrained operational contexts despite their lower accuracy ceiling.

4 Discussion

4.1. Contributions

This study provides a systematic comparison of Time Series Foundation Models for residential energy demand forecasting at 15-minute granularity, offering three main contributions: (1) **empirical evidence that TSFMs achieve state-of-the-art accuracy with minimal feature engineering**, contrarily to classical ML models which require extensive domain-specific preprocessing; (2) **quantification of the accuracy–latency trade-off** across model scales, from 20M to 200M parameters; and (3) **demonstration that zero-shot TSFMs can match, and in some configurations even outperform, fine-tuned classical ML** without a significant increase in inference time, even compared to deep learning methods. [11]

4.2. Implications for DSO Deployment

These findings carry practical implications for DSO practitioners selecting forecasting tools. Zero-shot TSFMs especially lightweight variants such as Chronos-Bolt or Chronos-Mini which enable high-quality forecasting with no training data requirement and minimal infrastructure, making them especially attractive for new deployments or sparse-data settings. When historical data is available, fine-tuning a small foundation model (Chronos-Small, 46M parameters) represents the most accurate configuration identified in this benchmark, while remaining computationally accessible without dedicated GPU hardware.

Language model-based approaches (LLMTime) are not recommended for operational use given their inference latency and lack of accuracy advantage over purpose-built TSFMs. GAN-based models offer no measurable benefit over simpler alternatives in this setting.

4.3. Limitations and Future Work

The benchmark is currently limited to 11 households from a single geographic region (Konstanz). Future work will extend the benchmark to larger, more diverse residential datasets (e.g., UK-DALE, Irish CER Smart Metering Trial) and will investigate probabilistic forecasting capabilities of TSFMs, which are particularly relevant for uncertainty-aware DSO planning. The integration of exogenous covariates (weather forecasts, calendar features) into TSFM conditioning will also be explored, as several foundation models now natively support such inputs.

5 Conclusion

This paper presented a comprehensive benchmark of 12 forecasting approaches for day-ahead residential electricity demand forecasting at 15-minute resolution. Foundation models demonstrate that zero-shot TSFMs can match or exceed fine-tuned classical ML performance without task-specific training, providing the best accuracy–latency trade-off across model scales. Chronos-Small (fine-tuned) achieves the best overall accuracy (NRMSE 0.58 ± 0.14, MAPE 44.9 ± 12.3%), while lightweight zero-shot variants (Chronos-Bolt, NRMSE 0.66) offer strong performance with minimal computational overhead, making them well suited for operational DSO deployment at scale. The high absolute error levels, consistent with the individual-meter forecasting literature, reflect the inherent stochasticity of residential consumption and reinforce the need for probabilistic rather than purely point-based forecasting approaches. These results suggest that lightweight foundation models (20–200M parameters) represent a compelling new alternative for residential energy forecasting, combining strong generalisation with practical computational requirements.

Acknowledgements

The authors thank the OpenPowerSystemData project for providing open access to the residential consumption dataset used in this study.

References

[1] Pinson, P. et al. (2014). Benefits and challenges of electrical demand response: A critical review. Renewable and Sustainable Energy Reviews, 39, 686–699.

[2] Haben, S. et al. (2014). A new error measure for forecasts of household-level, high resolution electrical energy consumption. International Journal of Forecasting, 32(4), 1368–1379

[3] Ansari, A. F. et al. (2024). Chronos: Learning the language of time series. *arXiv:2403.07815*.

[4] Das, A. et al. (2023). A decoder-only foundation model for time-series forecasting. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[5] Gruver, N. et al. (2023). Large Language Models are Zero-Shot Time Series Forecasters. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[6] Siano, P. (2014). Demand response and smart grids — A survey. Renewable and Sustainable Energy Reviews, 30, 461–478.

[7] Gajowniczek, K. et al. (2017). Electricity forecasting on the individual household level enhanced based on activity patterns.PLoS ONE 12(4): e0174098.

[8] Wu, H. et al. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[9] Hong, T., Fan, S. (2016). Probabilistic electric load forecasting: A tutorial review. International Journal of Forecasting, 32(3), 914–938.

[10] Misiurek, K. et al. (2025). Review of Methods and Models for Forecasting Electricity Consumption. *Energies*, 18, 4032.

[11] Kong, W. et al. (2017). Short-Term Residential Load Forecasting Based on LSTM Recurrent Neural Network. IEEE Transactions on Smart Grid, 10(1), 841–851.