REGIME-AWARE CROSS-MARKET ELECTRICITY PRICE FORECASTING WITH TIMESFM AND LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Electricity price forecasting is challenging because electricity markets are influenced by complex and rapidly changing interactions among demand, renewable generation, weather conditions, transmission constraints, market rules, and operational events. Time-series foundation models such as TimesFM provide a promising backbone for modeling numerical temporal dynamics, but they primarily rely on historical observations and may struggle when market regimes change. Large language models offer complementary capabilities for interpreting unstructured information, including regulatory updates, market announcements, operational reports, and extreme-event descriptions, but they are not designed to directly produce calibrated numerical forecasts. We propose a regime-aware cross-market electricity price forecasting framework that combines TimesFM with large language models. TimesFM models the temporal dynamics of electricity prices and related numerical variables, while a large language model converts time-stamped textual information into structured market-regime representations. These representations are integrated through a lightweight market-adaptation module, allowing a shared forecasting backbone to capture common patterns across markets while adapting to market-specific characteristics. The framework further supports probabilistic forecasting and explicitly models price spikes and uncertainty associated with regime transitions. We design a leakage-controlled evaluation protocol and conduct cross-market transfer experiments to assess the model's robustness under distribution shifts and limited data availability. The study evaluates forecasting accuracy, probabilistic calibration, extreme-price prediction, transferability, and computational efficiency, providing a systematic investigation of how language-derived market knowledge can enhance time-series foundation models for electricity price forecasting.

1 INTRODUCTION

Electricity prices change for reasons that are only partly visible in the price series itself. A sudden transmission constraint, a change in market rules, an unexpected renewable-output shortfall, or an operational announcement can alter the relationship between load, generation, and price within a few hours. This makes the problem different from forecasting a stationary seasonal signal. It also creates a practical asymmetry: numerical histories tell us what the market has been doing, while contemporaneous documents can describe why the relationship may be changing.

Modern time-series foundation models provide a strong starting point for the numerical part of this problem. TimesFM is a decoder-only foundation model trained for forecasting across many time-series domains (Das et al., 2024). Its strength is useful here: a shared backbone can transfer recurring temporal structure across markets without requiring a separate model for every province. Yet a numerical backbone has limited access to information that first appears in text. Large language models (LLMs) can parse such material, but direct numerical generation is poorly matched to calibrated electricity-price forecasting. We therefore use the LLM as an information extraction layer, not as the price forecaster.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

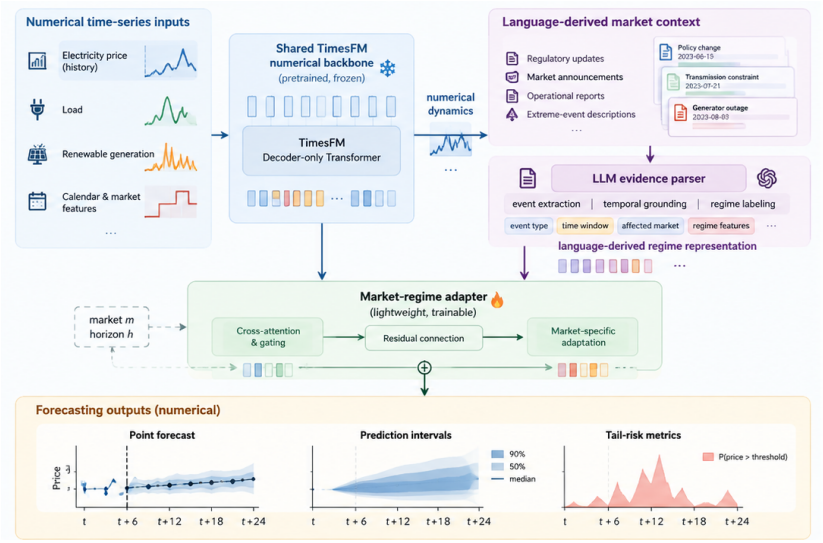


Figure 1: Regime-aware cross-market forecasting framework. TimesFM models the numerical dynamics, while the language model converts time-stamped official documents into structured regime evidence. The market-regime adapter combines the two streams only after origin-time alignment and produces point, interval, and tail-risk forecasts.

This paper asks a narrower question than whether an LLM can replace a forecasting model: *when does language-derived market context provide incremental value beyond a strong numerical backbone?* The distinction matters. Aggregate MAE can conceal a useful event-window signal, while an apparent gain can also arise from publication timing, missing-document patterns, or an unaligned feature that leaks future information.

Our framework, REGIME-TIMESFM, combines a shared TimesFM forecast with a residual market-regime adapter. The adapter receives numerical state variables, market identity, forecast horizon, and structured text representations derived from official market documents. Every text feature is aligned to the forecast origin. We evaluate both a strict observed-time protocol and a retrospective publication-proxy protocol, retain failed extractions as explicit missingness, and freeze one primary metric and one primary comparison before the final reporting pass.

The empirical result is deliberately more qualified than a universal-improvement claim. TimesFM and numerical covariates are the most reliable components. Language-derived context yields localized gains, especially in selected medium- and long-horizon regime strata, but it does not improve every market or every spike window. We view this pattern as a result rather than a failure: it identifies the conditions under which text is useful and prevents the paper from attributing every improvement to the language module.

Our contributions are fourfold:

- We formulate a regime-aware cross-market forecasting model in which TimesFM supplies the shared numerical backbone and an LLM supplies time-aligned market context.
- We provide a leakage-controlled protocol with strict text availability, publication-proxy sensitivity, frozen model comparison, explicit missing-text handling, and external numerical holdout evaluation.
- We report regime-conditional analyses and time-valid text controls that separate document availability, lexical content, policy/event structure, and prior-donor context.
- We release a reproducible result package with prediction files, manifests, hashes, the 87-article DeepSeek processing record, and a transparent boundary on what the external holdout does and does not establish.

2 Related Work

Electricity price forecasting. Electricity price forecasting has long combined autoregressive structure, calendar effects, exogenous variables, and probabilistic methods (Weron, 2014; Nowotarski & Weron, 2018). Day-ahead forecasting studies have shown that price spikes and changing market conditions can dominate average performance (Ziel & Weron, 2018). Our focus is complementary: we study whether official text provides a measurable regime signal after a foundation model has already captured numerical dynamics.

Time-series foundation models. Neural forecasting models such as DeepAR, N-BEATS, and the Temporal Fusion Transformer established strong probabilistic and multi-horizon baselines (Salinas et al., 2020; Oreshkin et al., 2020; Lim et al., 2021). Recent foundation models extend this idea across datasets and domains. TimesFM is central to our study because it offers a pretrained numerical backbone without requiring us to attribute numerical gains to a newly trained architecture (Das et al., 2024).

Language models for time series. Several studies explore prompting or adapting language models for temporal data (Gruver et al., 2023; Jin et al., 2024; Nie et al., 2023). These approaches establish that language models can encode useful temporal patterns, while also exposing the limits of treating numerical forecasting as ordinary text generation. We take a different division of labor: the LLM structures external documents, and TimesFM remains responsible for numerical forecasting.

Evaluation under temporal shift. Time-series evaluation must respect ordering and information availability (Bergmeir et al., 2018). Probabilistic evaluation benefits from proper scoring rules and calibration diagnostics (Gneiting & Raftery, 2007). We combine these principles with paired comparisons, tail metrics, missing-text sensitivity, and an external later-period numerical holdout.

3 Problem Formulation

For market m at forecast origin t, let $x_{\leq t,m}$ denote the observed numerical history, $z^{\text{num}}_{t,m}$ the engineered numerical state, and $z^{\text{text}}_{t,m}$ the structured representation of documents available by t. For horizon h, the forecast is

$$\widehat{y}_{t:t+h,m} = f_{\text{TimesFM}}(x_{\leq t,m}) + g_{\theta,h}\left(z^{\text{text}}_{t,m}, z^{\text{num}}_{t,m}, m\right). \tag{1}$$

The first term is shared across target markets. The second is a residual adapter fit separately for each evaluated horizon. The target is the hourly real-time or spot price at $t + h$; all prices and physical variables are aligned to the same hourly grid.

A document d is eligible at origin t only if its availability timestamp $a(d)$ satisfies $a(d) \leq t$. In the strict protocol, $a(d)$ is the first observed or publication timestamp available in the frozen packet. The publication-proxy protocol uses a fixed 24-hour release buffer and is reported as a retrospective sensitivity analysis. Text is retained for a finite memory window, 168 or 720 hours. Failed extraction is a separate missing indicator, not an empty event.

3.1 Formal properties

The following statements describe protocol properties, not performance guarantees.

Proposition 1 (origin-measurability). If text features use only documents with $a(d) \leq t$ and physical features are shifted by their declared lag ℓ, then $z^{\text{text}}_{t,m}$ and $z^{\text{num}}_{t,m}$ are measurable with respect to information available at t.

Proof sketch. The text representation is a finite sum over $\mathcal{D}_t = \{d : a(d) \leq t < t+W\}$, so it uses only origin-time records. The physical representation indexes the covariate grid at or before $t - \ell$. Concatenation, bounded interactions, and a deterministic fitted adapter preserve measurability. This property is exactly why the implementation audits origin-time alignment.

Proposition 2 (residual decomposition). For backbone forecast $b_{t,h}$ and adapter correction $\delta_{t,h}$,

$$y_{t+h} - \widehat{y}_{t+h} = (y_{t+h} - b_{t,h}) - \delta_{t,h}.$$

The language stream can reduce paired absolute error only when its correction has the appropriate sign and magnitude.

Proof. Substitute $\widehat{y}_{t+h} = b_{t,h} + \delta_{t,h}$ and rearrange. The identity motivates origin-matched comparisons, but it does not imply that an estimated correction is beneficial.

Proposition 3 (conditional value decomposition). For regime stratum R and paired absolute errors e^N, e^L,

$$\mathbb{E}[e^L - e^N] = \Pr(R)\mathbb{E}[e^L - e^N \mid R] + \Pr(R^c)\mathbb{E}[e^L - e^N \mid R^c].$$

Therefore a useful event-window effect can disappear in pooled MAE when event windows are rare or the effect reverses outside them.

Proof. Apply the law of total expectation to the paired error difference. This identity motivates the regime-conditional tables in the main paper.

4 METHOD

4.1 NUMERICAL BACKBONE AND STATE FEATURES

TimesFM produces a 24-step point forecast from the recent price history. We use its horizon-specific output as the backbone term in Eq. 1. Numerical features include price lags, rolling means and standard deviations, calendar harmonics, and delayed load and renewable-generation covariates. The physical covariates are lagged by 24 hours in the primary protocol. This conservative lag prevents the adapter from using a variable before its assumed availability.

4.2 TEXT-TO-REGIME REPRESENTATION

The LLM processes only the frozen set of public official articles. It extracts structured event classes, policy-state fields, market attribution, time evidence, and source spans. The resulting representation contains three groups: event-class counts, policy-state features, and document/lexical controls. The LLM never receives price observations, test labels, forecasts, or errors, and it never produces the final price forecast.

For each origin, event counts are accumulated over the memory window. If an article is unavailable or its extraction fails, its semantic fields contribute no event count and its missingness contributes an explicit control. This choice preserves the distinction between "no extracted event" and "failed extraction."

4.3 MARKET-REGIME ADAPTER

The adapter is a residual Ridge model with bounded state interactions:

$$g_{\theta,h}(u, v, m) = \rho_h \, \mathrm{Ridge}_{\theta_h}\left([u, v, \tanh(u_S) \otimes v, \, e_m]\right), \tag{2}$$

where $u = z^{\mathrm{num}}$, $v = z^{\mathrm{text}}$, S indexes numerical state coordinates, e_m is the market indicator, and ρ_h is a shrinkage coefficient. The regularization and shrinkage are chosen on the tune split only. The adapter is lightweight and horizon-specific. It corrects the TimesFM forecast instead of replacing it.

4.4 PROBABILISTIC AND TAIL OUTPUTS

We calibrate residual quantiles on a separate calibration split to produce 90% intervals. Auxiliary logistic heads estimate high-price and low-price risk. The high-price label is defined from a training-only threshold for each market and split. We report MAE, RMSE, high-price MAE, Brier score, coverage, interval width, and interval score. These heads are evaluation components, not claims that TimesFM itself provides native calibrated probabilities.

Table 1: Algorithmic roles and information boundaries.

Step	Operation	Information boundary
1	Forecast backbone	Historical numerical observations through origin t
2	Text extraction	Public document fields only; no prices or labels
3	Availability filter	Keep d iff $a(d) \leq t$
4	Adapter fitting	Train/tune splits; horizon-specific residual
5	Calibration	Calibration split for intervals and tail heads
6	Reporting	Test and holdout labels used once

5 Evaluation Protocol

5.1 Data and frozen analysis

The main study uses Guangdong and Jiangsu hourly prices with aligned physical covariates and 76 accepted text records. Fourteen failed records remain explicitly missing, while 13 deferred or quarantined records are excluded. The full numerical-context run contains 114 cases, 1,926 model-selection rows, 5,676 metric rows, and 114 prediction files. No forecasting experiment makes a provider call.

The primary metric is test MAE. The primary comparison is `llm_conditioned` versus `numeric_physical`. This comparison tests incremental language-conditioned adaptation beyond the strongest predeclared numerical covariate adapter. TimesFM remains the backbone reference. We report secondary metrics and paired bootstrap summaries as exploratory.

5.2 Cross-market transfer

We evaluate local models and transfer protocols with 0, 1, 7, and 30 days of target-market adaptation. Splits are chronological and enforce an embargo between train, tuning, calibration, and test origins. No test error or external-holdout label is used for model choice.

5.3 Regime strata and text controls

We stratify test origins by text-event exposure, volatility, price spikes, regime transitions, policy intensity, event intensity, and news intensity. Strata are descriptive. Spike labels use outcomes only for evaluation, while origin-time features use data available by the origin. For text controls, we compare aligned structured text, document controls, lexical controls, policy/event features, strictly earlier prior-donor controls, and strict observed-time alignment.

5.4 External evaluation status

We withhold the earlier Sichuan numerical scores from this manuscript. A post-run audit found inconsistent timestamp conventions between the candidate extension and the cached numerical grid, so those predictions cannot support an untouched-holdout claim. A corrected external evaluation remains outstanding.

6 Results

6.1 External numerical holdout

The external score table is intentionally omitted. The study reports no external numerical ranking until timestamps are reconciled and TimesFM inference is rerun from the corrected grid.

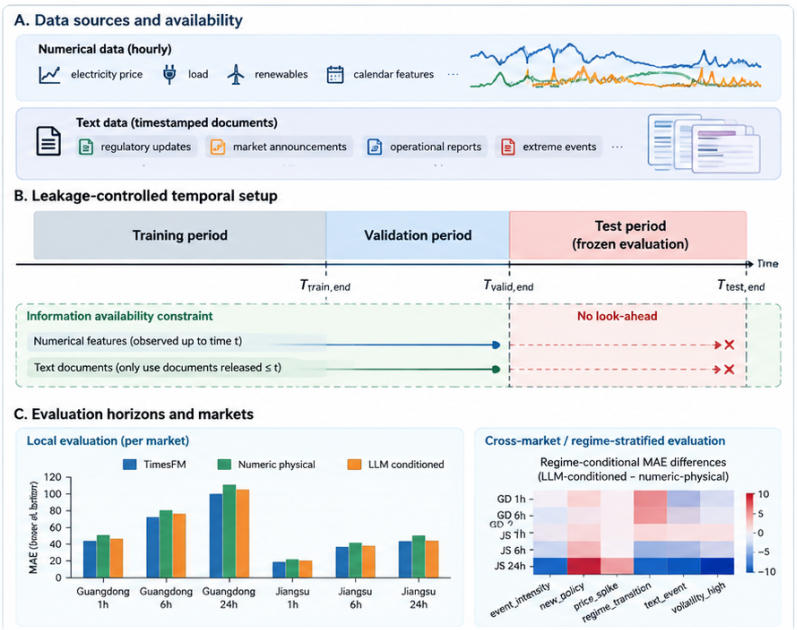


Figure 2: Leakage-controlled evaluation protocol. Every document and numerical covariate is filtered by its availability time before model fitting; test labels are accessed only once for final evaluation.

Table 2: External evaluation audit. Numeric scores are withheld pending timestamp reconciliation.

Audit item	Status
Candidate period	Available
Aligned external text	Missing
Timestamp convention	Inconsistent
Cached input grid	Unreconciled
Numerical score claim	Withheld
Corrected untouched test	Outstanding

6.2 Primary comparison

The table gives a useful warning against a single headline number. The language-conditioned model is close to the numerical adapter in Guangdong and improves on it in Jiangsu at 24 hours, but it is worse at Jiangsu 1 hour. This is consistent with the intended role of the text stream: it describes slower regime information more naturally than it resolves every next-hour price movement.

6.3 When does text help?

Table 5 summarizes the central result. Language-derived context is most useful in selected medium- and long-horizon strata in Jiangsu. It does not help in every event or spike window. The result supports a conditional regime interpretation, not a universal text advantage.

6.4 Alignment ablations

The alignment study compares correctly aligned structured text with document-only controls, lexical controls, policy/event features, prior-donor controls, and strict observed-time alignment. In Jiangsu at 6 hours, aligned text obtains MAE 31.38, compared with 33.56 for lexical context and 32.80–32.80 for two prior-donor controls. In Guangdong at 1 hour, all variants are close. These results suggest that the useful signal, when present, is tied to structured and time-aligned context rather than simply adding arbitrary text-derived dimensions.

Table 3: Primary local comparison under the frozen protocol. MAE is the primary metric.

Market	Horizon	TimesFM	Num.-physical	LLM-conditioned
Guangdong	1h	57.36	56.51	56.74
Guangdong	6h	103.50	102.80	102.80
Guangdong	24h	114.49	114.49	114.49
Jiangsu	1h	13.41	13.72	14.30
Jiangsu	6h	31.17	31.46	31.38
Jiangsu	24h	33.18	42.04	34.33

Table 4: Secondary metrics for the six-hour local evaluation. These values are descriptive and do not define model selection.

Market	Model	RMSE	High MAE	Coverage	High Brier
GD	TimesFM	159.429	230.294	.867	.117
GD	Numeric	157.851	222.821	.863	.153
GD	Language	157.996	223.664	.868	.155
JS	TimesFM	61.016	92.532	.820	.059
JS	Numeric	60.913	92.311	.826	.073
JS	Language	60.637	93.337	.826	.067

7 Discussion

The experiments support a modest claim. TimesFM is the main forecasting engine. Text is an auxiliary view of market state. It helps when a document describes a regime whose effect persists beyond the immediate next hour, and it helps less when the target is dominated by high-frequency noise or a price spike that is not explicitly documented. The cross-market setting makes this distinction visible because the same numerical backbone encounters different market mechanisms and different document coverage.

The external holdout is not used to support a numerical ranking in this version. This boundary is intentional: a corrected evaluation requires timestamp reconciliation and fresh inference before labels are inspected.

8 Limitations and Broader Impact

The text packet is small relative to the number of hourly price observations, and 14 articles remain missing after validation. The missingness analysis is a sensitivity analysis rather than a causal estimate. The candidate Sichuan holdout is withheld after an audit found timestamp and input-grid inconsistencies; it is not used as evidence until fresh inference is rerun. Event labels extracted from public articles are not a semantic gold standard. The publication-proxy protocol is retrospective and must be separated from the strict observed-time analysis.

Forecasts may affect bidding, scheduling, and risk decisions. The model should therefore be used as decision support with human oversight. A language-derived signal should never override market-safety controls or be interpreted as causal evidence about a policy or operational event.

9 Reproducibility and AI Disclosure

The forecasting experiments use zero provider calls. DeepSeek processed a fixed authorized batch of 87 public official articles. The returned model was `deepseek-flash`. Seventy-three records passed schema validation and 14 were retained as missing. Reported usage was 114,512 prompt tokens and 15,815 completion tokens, 130,327 total. No price observations, test labels, predictions, or private documents were sent to the provider. The complete disclosure and hashes are included in the supplementary material.

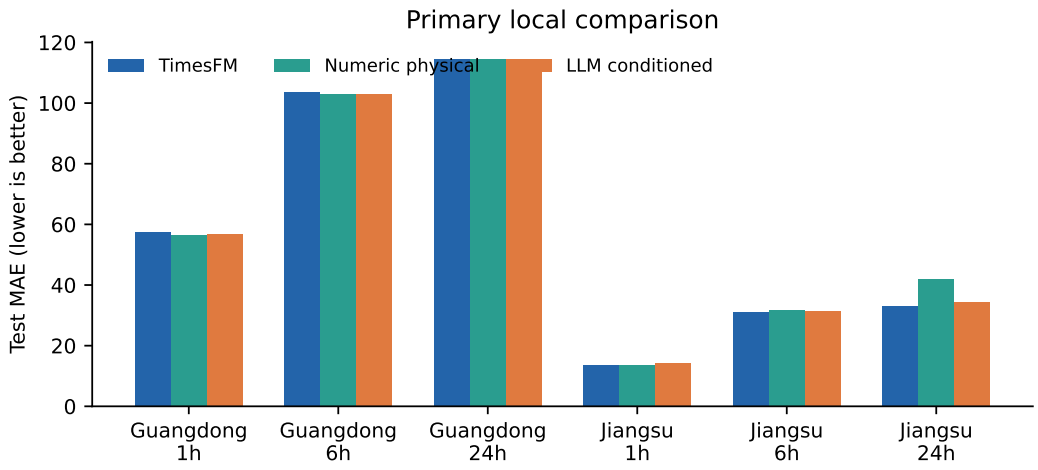


Figure 3: Primary local MAE comparison under the frozen protocol.

Table 5: Selected regime-conditional MAE differences, LLM-conditioned minus numeric-physical. Negative values favor the language-conditioned model.

Market	Horizon	Stratum	n	ΔMAE
GD	1h	text event	28	+1.45
GD	1h	price spike	136	+0.74
JS	6h	text event	116	-0.50
JS	6h	transition	72	-0.73
JS	24h	transition	73	-12.17
JS	24h	high volatility	85	-8.40

10 Conclusion

We presented a regime-aware cross-market forecasting framework that assigns different jobs to numerical and language models. TimesFM captures numerical dynamics and remains the strongest, most reliable component. An LLM can add useful market context, but its value is conditional on market, horizon, and regime. The main contribution is therefore an auditable way to test language-derived context under temporal leakage constraints, rather than a claim that language features improve every forecast.

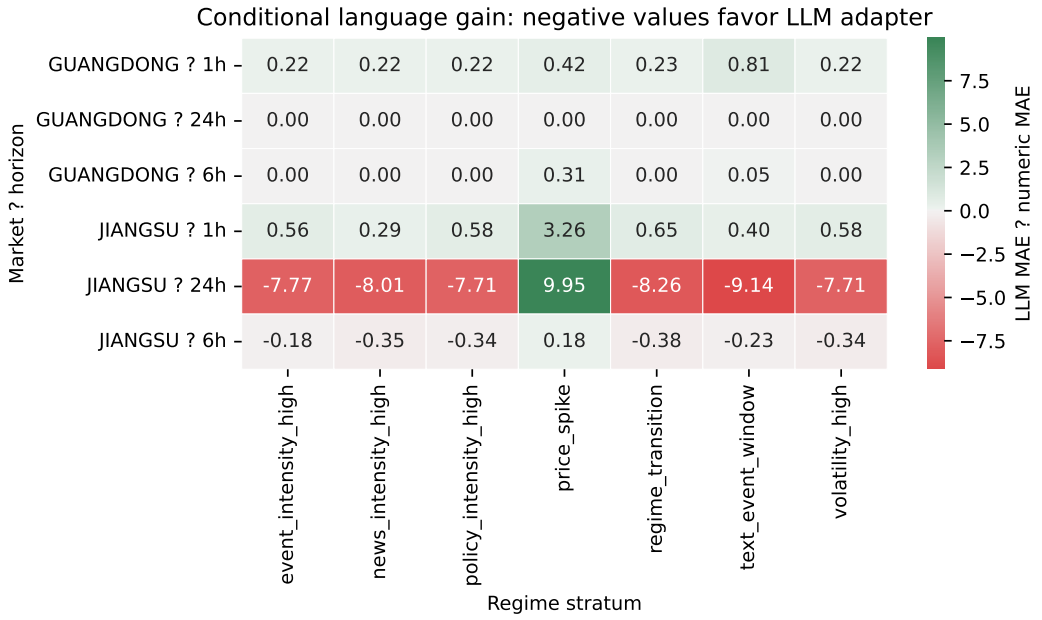


Figure 4: Regime-conditional MAE differences. Negative values favor the language-conditioned adapter.

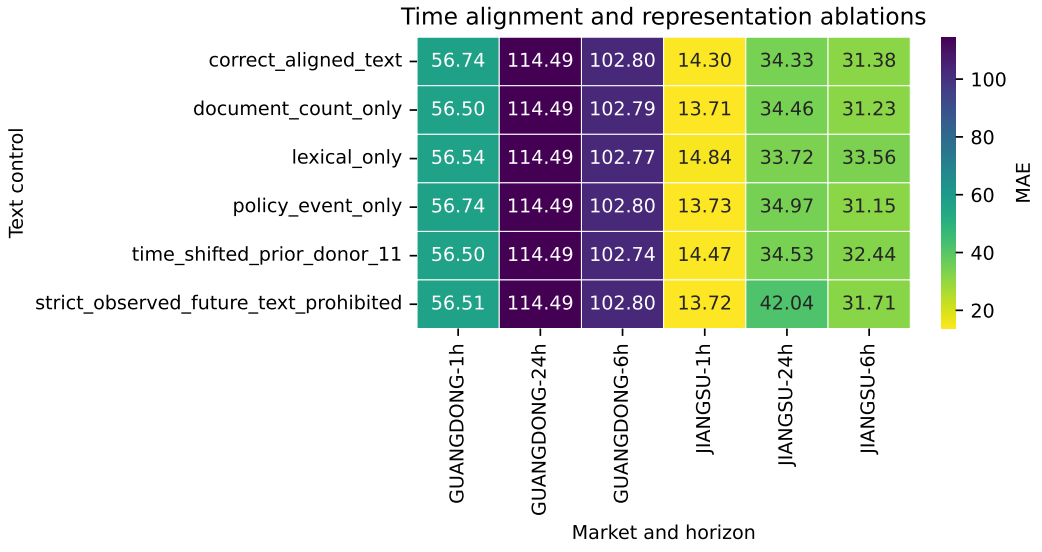


Figure 5: Text representation and temporal alignment controls.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Christoph Bergmeir, Rob J. Hyndman, and Bonsoo Koo. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83, 2018. doi: 10.1016/j.csda.2017.11.003.

George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. Wiley, 5th edition, 2015.

Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Antonio J. Conejo, Javier Contreras, Ramon Espinola, and Antonio B. Plazas. Forecasting electricity prices for a day-ahead pool-based electric energy market. *International Journal of Forecasting*, 21(3):435–462, 2005. doi: 10.1016/j.ijforecast.2004.12.005.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 10525–10542. PMLR, 2024.

Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995. doi: 10.1080/07350015.1995.10524599.

Jesse Dodge et al. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 1286–1305, 2022.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Tilmann Gneiting, Fadoua Balabdaouze, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, 69(2):243–268, 2007. doi: 10.1111/j.1467-9868.2007.00587.x.

Nate Gruver et al. Large language models are zero-shot time series forecasters. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Joel Hestness et al. Deep learning scaling is predictable, empirically, 2017. arXiv:1712.00409.

Hansika Hewamalage, Christoph Bergmeir, and Kasun Bandara. Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, 37 (1):388–427, 2021. doi: 10.1016/j.ijforecast.2020.06.008.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 3rd edition, 2021. URL `https://otexts.com/fpp3/`.

Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006. doi: 10.1016/j.ijforecast.2006.03.001.

Gautier Izacard et al. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43, 2023.

Ming Jin et al. Time-llm: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Bryan Lim, Sercan O. Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4): 1748–1764, 2021. doi: 10.1016/j.ijforecast.2021.03.012.

Katarzyna Maciejowska, Jakub Nowotarski, and Rafał Weron. Probabilistic forecasting of electricity spot prices using factor models. *Energy Economics*, 58:236–254, 2016. doi: 10.1016/j.eneco.2016.06.012.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4):802–808, 2018. doi: 10.1016/j.ijforecast.2018.06.001.

Pablo Montero-Manso and Rob J. Hyndman. Principles and algorithms for forecasting groups of time series: Locality and globality. *International Journal of Forecasting*, 37(4):1632–1653, 2021. doi: 10.1016/j.ijforecast.2021.03.004.

Yuqi Nie et al. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Jakub Nowotarski and Rafał Weron. Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renewable and Sustainable Energy Reviews*, 81:1548–1568, 2018. doi: 10.1016/j.rser.2017.05.234.

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you? explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016. doi: 10.1145/2939672.2939778.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020.

Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32, 2019.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020. doi: 10.1016/j.ijforecast.2019.07.001.

Leonard J. Tashman. Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, 16(4):437–450, 2000. doi: 10.1016/S0169-2070(00)00065-0.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 10014–10037, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Rafał Weron. Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, 30(4):1030–1081, 2014. doi: 10.1016/j.ijforecast.2014.08.008.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

Gerald Woo, Chenghao Liu, Doyen Sahoo, and Akshat Kumar. Etsformer: Exponential smoothing transformers for time-series forecasting, 2022. arXiv:2202.01381.

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*, volume 34, 2021.

Haixu Wu, Tengge Hu, Yong Liu, Hu Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 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, pp. 11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317.

Yunhao Zhang et al. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations*, 2023.

Haoyi Zhou, Shanghang Zhang, Jie men Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021. doi: 10.1609/aaai.v35i12.17325.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, volume 162, pp. 27268–27286. PMLR, 2022.

Florian Ziel and Rafał Weron. Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks. *Energy Economics*, 70:295–308, 2018. doi: 10.1016/j.eneco.2017.12.016.