BRIDGING THE EXPERIENCE GAP: SEMANTIC RETRIEVAL AND CONDITIONAL RESIDUAL TRANSFER FOR TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Time series forecasting draws on historical observations, but abundant historical data does not necessarily provide relevant experience for a particular forecast. An entity may face conditions absent from its own history, while useful experience exists across other entities and contexts. Drawing on this experience presents two challenges: 1) identifying relevant historical cases through semantic connections rather than numerical similarity; 2) adapting their responses across differences in timing, magnitude, and shape to improve the target forecast. To address these challenges, we propose **SCoRe**, a plug-and-play framework for augmenting frozen forecasting models. It retrieves semantically analogous historical cases and adapts their forecast residuals to correct the target forecast. **For semantic retrieval**, a large language model jointly characterizes entity attributes, historical states, and known future conditions, exposing contextual connections that guide the retrieval of potentially transferable residuals. **For response adaptation**, the retrieved residuals are aligned with the target conditions. Learned directional gain estimates and candidate interactions then guide the joint selection and weighting of residual corrections. By combining semantic retrieval with explicit response adaptation and joint utility assessment, the framework enables forecast correction when direct historical matches are unavailable. Our evaluation examines forecasting accuracy, transfer across entities and contexts, and negative transfer across multiple forecasting backbones. Code is available at https://anonymous.4open.science/r/TextRAG-F67B.

1 INTRODUCTION

Time series forecasting is widely used in retail forecasting (Fildes et al., 2022), energy scheduling (Hong & Fan, 2016) and transportation operations (Vlahogianni et al., 2014) to inform downstream decisions. In practice, however, forecasts often need to be made for target entities that lack direct historical experience under the conditions they are about to encounter: existing observations may cover their routine operating states but not the forthcoming external conditions. Such gaps in contextual coverage are also prevalent in large-scale retail data. Our analysis of 87 million real-world sales forecasting samples from a major e-commerce platform reveals that, during key promotional festivals, approximately 45%–50% of products have no historical records for the same festival at the time of prediction. This finding indicates that even with massive amounts of observations, direct experience for a target product under a particular event may remain highly limited. Similar gaps may arise in other domains: a transportation region may lack observations during a particular type of large-scale event, or an energy consumer may lack load records under a certain type of extreme weather. These settings share a common problem: direct historical observations are unavailable for the combination of the target entity and its future context, while potentially useful experience is scattered across other entities or related contexts.

As illustrated in Figure 1, effectively exploiting historical experience across entities and contexts presents two interrelated challenges. **First, how can historical cases with potential transfer value be identified across entities and contexts when direct history is limited?** This identification is crucial because the most relevant reference cases may span different entities and covariate conditions, linked by shared demand drivers rather than observable similarity. For a rose gift box facing

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.

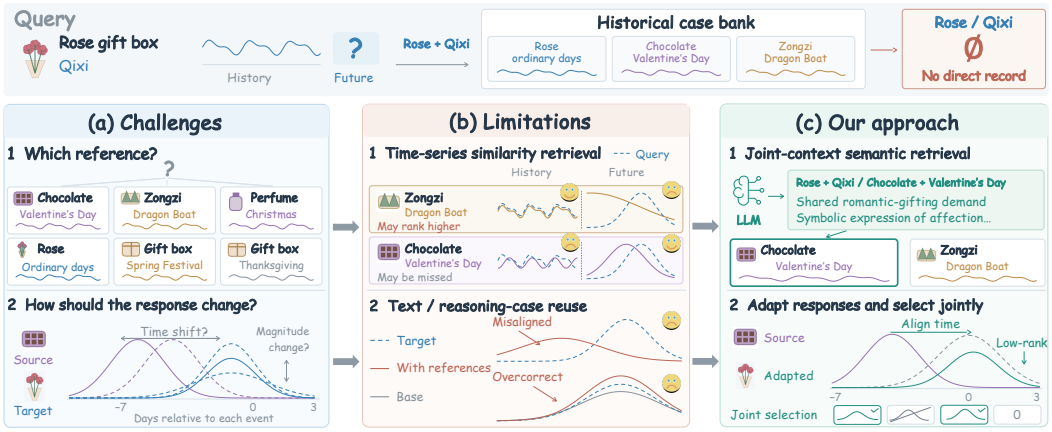


Figure 1: Forecasting with limited entity–context history. **(a) Challenges:** discovering potentially transferable cases across entities and contexts and adapting their responses. **(b) Existing limitations:** trajectory similarity may miss relevant cases; semantic relevance does not guarantee predictive utility. **(c) Our approach:** LLM-based semantic retrieval discovers candidates for conditional adaptation and joint correction selection.

its first Qixi promotion, Dragon Boat gift boxes may have similar sales histories but serve a different gifting purpose, whereas Valentine's Day chocolates may offer a relevant analogue because both express romantic affection, despite dissimilar sales histories. Recognizing this potential connection requires considering how product attributes and event meanings jointly shape the purchasing context. One class of existing methods retrieves reference cases based on historical trajectories, entity context, or covariate relationships and incorporates the retrieved numerical information into forecasting (Han et al., 2025; Ning et al., 2026; Liang et al., 2026; Tang et al., 2026). Although these methods exploit similarities in observed states and conditions, such matching may overlook potential analogies that are insufficiently represented in training observations. In particular, when festivals are encoded as discrete identifiers, numerical and identifier matching alone may miss the shared romantic-gifting context of Qixi roses and Valentine's Day chocolates.

Second, how can retrieved cases be converted into effective predictive evidence under target conditions? This transformation is necessary because a relevant historical case does not necessarily exhibit a response that can be directly transferred to the target (Wang et al., 2019; Zhang et al., 2022). Although chocolate and roses may share the same romantic gifting demand, their sales responses can differ substantially in timing: chocolate purchases may peak several days before Valentine's Day, whereas freshness constraints may concentrate rose purchases on Qixi itself. Differences in baseline sales and promotional intensity may further affect response magnitude, making direct reuse unreliable. Another class of existing methods reuse retrieved cases through reasoning guidance or learned fusion and routing (Wang et al., 2026; Tang et al., 2026; He et al., 2026). However, they lack an explicit connection between adapting source responses to differences in entity attributes and context and evaluating the joint predictive value of the adapted cases. Reference relevance alone cannot determine how to adapt source responses or whether combining them improves forecasts.

To address these challenges, we propose **SCoRe**, a framework for time series forecasting that combines **S**emantic retrieval with **Co**nditional **Re**sidual transfer. Implemented as a plug-and-play output correction module, SCoRe can augment different base forecasting models. **To discover potentially transferable references beyond numerical similarity**, we introduce joint context semantic retrieval, where an LLM jointly represents entity attributes, historical states, and known future conditions to identify relevant cases across different entities and contexts, together with their historical forecast residuals. **To convert these references into predictive evidence**, we conditionally transfer the retrieved residuals to the target context and determine their contributions according to estimated predictive utility. This allows SCoRe to exploit semantically relevant historical experience while suppressing unreliable or uninformative corrections. Extensive experiments across diverse benchmarks and forecasting backbones demonstrate the effectiveness and versatility of SCoRe, showing that semantic retrieval coupled with conditional residual transfer can improve forecasting accuracy. Our main contributions are as follows:

- We introduce joint context semantic retrieval that leverages LLM generated descriptions of entity attributes, historical states, and known future conditions to identify potentially transferable cases across different entities and contexts beyond observed numerical similarities.
- We develop conditional residual transfer with joint selection after adaptation, transforming retrieved residuals into corrections tailored to the target conditions and evaluating their joint predictive utility, while retaining the base forecast when the retrieved evidence is insufficient to avoid negative transfer.
- Extensive experiments across diverse benchmarks and forecasting backbones demonstrate that SCoRe consistently improves forecasting accuracy and effectively bridges the experience gap by leveraging relevant cases from other entities and contexts when the target history provides insufficient predictive evidence.

2 Related Work

Covariate-Aware Time Series Forecasting. Covariates provide information beyond the target history and commonly include static attributes, observed variables, and known future variables. Deep covariate-aware forecasters learn nonlinear dependencies between these signals and the target series (Salinas et al., 2020). TFT (Lim et al., 2021) integrates heterogeneous covariates through variable selection, gating, and temporal attention, while TiDE (Das et al., 2023) adopts an MLP-based encoder. Transformer-based methods such as TimeXer (Wang et al., 2024) and CITRAS (Yamaguchi et al., 2026) further model target–covariate interactions through specialized representations and attention mechanisms. Recent pretrained forecasters have extended covariate modeling toward more general settings (Woo et al., 2024; Dange & Sarawagi, 2025). Chronos-2 (Ansari et al., 2025) supports multivariate and covariate-informed zero-shot forecasting, while UniCA (Han et al., 2026) adapts frozen time-series foundation models to heterogeneous covariates through representation homogenization and modular fusion. Despite their effectiveness, these approaches primarily encode covariate effects through model parameters or in-context interactions. When a target entity lacks observations under a forthcoming context, they do not explicitly retrieve and reuse potentially transferable responses from other entities or related contexts.

Retrieval-Augmented and Semantic Time Series Forecasting. Retrieval-augmented forecasting complements parametric models with historical cases accessed at inference time. Rather than relying exclusively on information compressed into model parameters, these methods retrieve relevant training windows and reuse their subsequent observations or latent representations. RAFT (Han et al., 2025) retrieves historical trajectories similar to the query history and provides their future values as additional predictive signals. TS-RAG (Ning et al., 2026) retrieves time-series segments using a pretrained encoder and dynamically integrates them into a time-series foundation model through a learnable mixture-of-experts module. However, retrieval based on numerical similarity may overlook transferable cases whose relevance stems from shared entity roles or contextual meanings rather than similar trajectories (Williams et al., 2024). Language-assisted methods offer semantic cues (Wang et al., 2026; Xue & Salim, 2023), but their use as prompting or reasoning context does not explicitly establish which historical responses are transferable or how they should be adapted to target conditions. SCoRe addresses this gap by coupling joint-context semantic retrieval with conditional residual transfer and post-adaptation joint selection, connecting reference discovery to numerical adaptation and predictive utility.

3 Methodology

3.1 Problem Formulation and Overview

At forecast origin q, the available information is $\mathcal{I}_q = (\mathbf{X}_q, \mathbf{C}_q^{\text{hist}}, \mathbf{C}_q^{\text{known}}, \mathbf{a}_q)$: target history $\mathbf{X}_q \in \mathbb{R}^L$, historical covariates, known future covariates, and optional entity attributes. Future inputs must be available at q, not realized future measurements. A frozen backbone produces $\mathbf{B}_q = f_\phi(\mathcal{I}_q) \in \mathbb{R}^H$ for target $\mathbf{Y}_q$, using the inputs supported by its architecture. SCoRe learns an additive correction to the forecast output, without modifying backbone representations.

SCoRe couples semantic retrieval with conditional residual transfer to exploit historical experience across entities and contexts. As shown in Figure 2, it comprises four modules. *Semantic Residual*

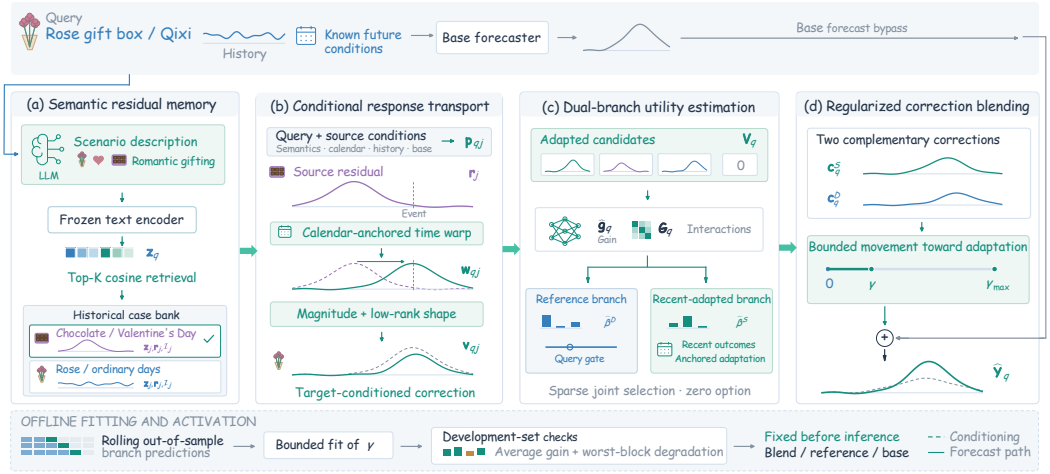


Figure 2: **Overview of SCoRe.** (a) Semantic Residual Memory (SRM) retrieves historical residuals through joint-context semantic descriptions. (b) Conditional Response Transport (CRT) aligns and adapts retrieved residuals to target conditions. (c) Dual-branch Utility Estimation (DUE) composes candidate corrections based on their joint predictive utility. (d) Regularized Correction Blending (RCB) combines the branch corrections to enhance the base forecast.

Memory (SRM) retrieves backbone-specific residuals using LLM-generated descriptions of entity attributes, historical states, and known future conditions. *Conditional Response Transport (CRT)* temporally aligns and adapts retrieved residuals to the query. *Dual-branch Utility Estimation (DUE)* combines adapted candidates based on their predicted utility and interactions through historical and recently adapted branches. *Regularized Correction Blending (RCB)* blends the branch corrections using a bounded coefficient fitted on rolling-validation predictions, subject to a development-set risk check. The resulting correction is added to the base forecast, with a separate residual memory and enhancement parameters for each backbone.

3.2 Semantic Residual Memory

Similarity between target histories alone can miss relevance arising from covariate conditions. SRM therefore uses an LLM to describe each forecasting scenario jointly from its history, covariates, and available entity information:

$$T_q = \mathcal{G}_{\text{LLM}}\big(\mathcal{S}_x(\mathbf{X}_q), \mathcal{S}_c(\mathbf{C}_q^{\text{hist}}), \mathcal{S}_f(\mathbf{C}_q^{\text{known}}), \mathbf{a}_q, \mathcal{D}_q\big)\,, \tag{1}$$

where $\mathcal{S}_x, \mathcal{S}_c, \mathcal{S}_f$ summarize the corresponding inputs and $\mathcal{D}_q$ contains domain metadata. The structured description includes pattern features, a historical summary, a qualitative hypothesis about future trends, and a combined scenario summary. The hypothesis uses only available inputs and serves as a retrieval cue, not a numerical forecast (Appendix E.1).

From a historical text index $\mathcal{M}$, we retrieve $\mathcal{J}_q = \text{TopK}_{j\in\mathcal{M}}\, \text{sim}_{\text{text}}(T_q, T_j)$ using cosine similarity of description embeddings. Descriptions and candidate rankings are precomputed and fixed during enhancement training. For downstream conditioning, a frozen BGE encoder (Xiao et al., 2024) provides normalized embeddings $\mathbf{z}_i = \text{Enc}(T_i)/\|\,\text{Enc}(T_i)\|_2$. Each valid source supplies a normalized backbone residual:

$$\mathbf{r}_j = (\mathbf{Y}_j - \mathbf{B}_j^{\text{hist}})/s_j, \qquad \mathbf{e}_q = (\mathbf{Y}_q - \mathbf{B}_q)/s_q, \tag{2}$$

where $s_i > 0$ is a scale estimated from the historical window with a fixed historical floor. Historical forecasts $\mathbf{B}_j^{\text{hist}}$ follow a protocol in which fitting and selection for the task precede their forecast windows; changing the backbone requires regenerating its residuals. Only sources with fully observed target windows at q are valid. Invalid candidates are masked without replenishment, so the pool may be empty. This memory supplies errors unexplained by the backbone, while semantic similarity only proposes references: transport and utility estimation determine their contributions.

3.3 CONDITIONAL RESPONSE TRANSPORT

Semantically related scenarios can still differ in response phase, duration, amplitude, and shape. CRT resolves these mismatches by transporting each source residual before evaluating its utility. Pair conditions $\mathbf{p}_{qj}$ concatenate encoded query/source conditions $\mathbf{h}_q, \mathbf{h}_j$, their difference and product, and fixed pair features such as similarity and scale ratios. The encoders use text, known calendar features, historical statistics, and base forecasts. Matching calendar anchors supplies a prior $\boldsymbol{\tau}^0_{qj}$, repaired toward identity when necessary. CRT learns bounded offsets and applies linear interpolation:

$$\boldsymbol{\tau}_{qj} = \boldsymbol{\tau}^0_{qj} + \alpha^{\mathrm{warp}}_{qj}\mathbf{d}_\theta(\mathbf{p}_{qj}), \qquad \mathbf{v}_{qj} = (a_{qj}\mathbf{I} + \mathbf{U}\mathbf{C}_{qj}\mathbf{U}^\top)\mathbf{W}_{\boldsymbol{\tau}_{qj}}\mathbf{r}_j. \tag{3}$$

The largest feasible $\alpha^{\mathrm{warp}}_{qj} \in [0,1]$ preserves source support, bounded displacement, and positive bounded local slopes. Offsets vanish at the endpoints; missing anchors disable warping. The fixed orthonormal basis $\mathbf{U} \in \mathbb{R}^{H\times r}$ supports a shape increment of low rank without projecting away the full response. We constrain $a_{qj} \in (a_{\min}, a_{\max})$ and $\|\mathbf{C}_{qj}\|_F < \rho < a_{\min}$. Thus temporal transport handles timing mismatch, while scaling and the shape increment adapt amplitude and shape. Appendix B establishes temporal feasibility and adapter norm bounds.

3.4 DUAL-BRANCH UTILITY ESTIMATION

Semantic relevance does not imply error reduction, and separately scoring candidates ignores their overlapping contributions (Krogh & Vedelsby, 1994). DUE instead evaluates their joint utility after transport. Write $\mathbf{V}_q = [\mathbf{v}_{q1}, \ldots, \mathbf{v}_{qK}]$, $\langle \mathbf{u}, \mathbf{v}\rangle_H = \mathbf{u}^\top\mathbf{v}/H$, and $\Delta = \{\boldsymbol{\beta} \geq 0 : \mathbf{1}^\top\boldsymbol{\beta} \leq 1\}$. For $\boldsymbol{\beta} \in \Delta$, the realized error reduction is

$$\mathcal{U}_q(\boldsymbol{\beta}) = \|\mathbf{e}_q\|^2_H - \|\mathbf{e}_q - \mathbf{V}_q\boldsymbol{\beta}\|^2_H = 2\boldsymbol{\beta}^\top\mathbf{g}^*_q - \boldsymbol{\beta}^\top\mathbf{G}_q\boldsymbol{\beta}, \tag{4}$$

where $g^*_{qj} = \langle \mathbf{e}_q, \mathbf{v}_{qj}\rangle_H$ and $\mathbf{G}_q = \mathbf{V}_q^\top\mathbf{V}_q/H$. Networks predict $\widehat{\mathbf{g}}_q$ from pair conditions, transported responses, and pool geometry, supervised by fully observed historical $\mathbf{g}^*_q$. We compute $\mathbf{G}_q$ exactly and approximately maximize $\widehat{\mathcal{U}}_q(\boldsymbol{\beta}) = 2\boldsymbol{\beta}^\top\widehat{\mathbf{g}}_q - \boldsymbol{\beta}^\top\mathbf{G}_q\boldsymbol{\beta}$ over Δ with a fixed number of Frank–Wolfe iterations (Jaggi, 2013), initialized at zero. Nonpositive estimated utility yields zero correction; inference never requires the query's future residual.

Temporal distribution shifts have motivated adaptive forecasting and online ensembling methods (Wen et al., 2023; Du et al., 2021; Kim et al., 2025). Because utility relationships can shift, DUE maintains two branches sharing frozen transported candidates. Reference branch D uses a historical utility network and a query gate $\alpha_q \in [0,1]$. Branch S uses a separately pretrained utility network adapted on recent fully observed windows, with frozen historical normalization, feature clipping, and parameter anchoring:

$$\mathbf{c}^D_q = \alpha_q s_q \mathbf{V}_q\widehat{\boldsymbol{\beta}}^D_q, \qquad \mathbf{c}^S_q = s_q\mathbf{V}_q\widehat{\boldsymbol{\beta}}^S_q. \tag{5}$$

D shrinks the historical composition according to the query; S updates the relationship between conditions and utility. Appendix C gives the utility identity and solver properties.

3.5 REGULARIZED CORRECTION BLENDING

Recent adaptation can capture residual shift but can also amplify estimation error (Chen et al., 2024). RCB balances these effects using one regularized mixing coefficient shared across queries within each fitted model, rather than an additional network conditioned on the query. Let $\mathbf{d}_q = \mathbf{c}^S_q - \mathbf{c}^D_q$. The final forecast is

$$\widehat{\mathbf{Y}}_q = \mathbf{B}_q + \mathbf{c}^D_q + \bar{\gamma}\mathbf{d}_q, \qquad 0 \leq \bar{\gamma} \leq \gamma_{\max} = 0.5. \tag{6}$$

The branch corrections still depend on the query; only their mixing coefficient is shared. To fit the coefficient without using branch predictions on their fitting samples, we generate paired corrections on rolling validation folds, denoted OOF corrections (Cerqueira et al., 2020; Wolpert, 1992). In each of three chronological folds, S is adapted and a separate D query gate is fitted on the available prefix, then both predict later windows reserved for validation. Fitting labels must mature before the fold boundary, and validation target windows must end within the fold. For OOF queries $\mathcal{O}$, define $\widetilde{\mathbf{r}}_q = \mathbf{Y}_q - \mathbf{B}_q - \widetilde{\mathbf{c}}^D_q$ and $\widetilde{\mathbf{d}}_q = \widetilde{\mathbf{c}}^S_q - \widetilde{\mathbf{c}}^D_q$. We minimize

$$\mathcal{L}_{\mathrm{blend}}(\gamma) = \sum_{q\in\mathcal{O}}\left[\|\widetilde{\mathbf{r}}_q - \gamma\widetilde{\mathbf{d}}_q\|^2_H + \lambda_b\gamma^2\|\widetilde{\mathbf{d}}_q\|^2_H\right], \quad \gamma \in [0, \gamma_{\max}], \tag{7}$$

where $\lambda_b = 0.1$ shrinks movement away from D in proportion to the energy of branch disagreement. This scalar quadratic has the analytical solution

$$\widehat{\gamma} = \mathrm{clip}_{[0,\gamma_{\max}]} \frac{\sum_{q\in\mathcal{O}} \langle \widetilde{\mathbf{r}}_q, \widetilde{\mathbf{d}}_q \rangle_H}{(1+\lambda_b)\sum_{q\in\mathcal{O}} \|\widetilde{\mathbf{d}}_q\|_H^2}. \tag{8}$$

We set $\widehat{\gamma} = 0$ for negligible total disagreement energy. Appendix D proves the solution and bounds the resulting movement. On a later development interval, we compare the mixture with D using

$$R_{\mathrm{dev}}(\widehat{\gamma}) = \mathrm{Avg}_q\, \delta_q + \lambda_r \max\{0, \max_{b\in\mathcal{B}} \mathrm{Avg}_{q\in b}\, \delta_q\}, \tag{9}$$

where δ_q is the MSE of the mixture minus that of D, normalized by the fixed variance of training targets, $\lambda_r = 0.5$, and $\mathcal{B}$ comprises temporal blocks of 168 rows. We set $\bar{\gamma} = \widehat{\gamma}$ if $R_{\mathrm{dev}} < 0$, and $\bar{\gamma} = 0$ otherwise, reverting to D. D has its own check of development MSE against the backbone and contributes zero if that check fails. If the mixture is also disabled, the final forecast equals the backbone. These checks govern activation, not guarantees of improvement on unseen queries.

3.6 Training and Inference

Training respects temporal label availability. CRT first learns from individual residual reconstruction and an oracle combination loss used only for training and validation, then remains frozen. Utility networks learn a location $\widehat{g}_{qj}$ and positive scale σ_{qj} through

$$\mathcal{L}_g = \mathrm{Avg}_q \frac{1}{K_q} \sum_{j\ \mathrm{valid}} \left[|g^*_{qj} - \widehat{g}_{qj}|/\sigma_{qj} + \log \sigma_{qj} \right], \tag{10}$$

averaging valid candidates within each nonempty query (K_q candidates). The D query gate minimizes forecast MSE; S adaptation adds parameter anchoring. RCB fits $\widehat{\gamma}$ by Equation 8 on rolling OOF predictions and uses only the later development interval to decide activation. Complete target windows remain within their assigned temporal intervals. All model parameters, the shared mixing coefficient, and activation decisions are fixed before testing; inference never uses query targets or refits the coefficient. Appendix E provides the detailed protocol and configurations.

4 Experiments

4.1 Experimental setup

Datasets. We consider four public datasets for forecasting with covariates: Electricity Price Forecasting (Lago et al., 2021) (EPF), Electricity Demand Forecasting (Aguilar Madrid & Antonio, 2021) (EDF), Bike Sharing (Fanaee-T, 2013), and M5 (Makridakis et al., 2022). EPF comprises five market-specific subsets: Belgium (BE), Germany (DE), France (FR), Nord Pool (NP), and Pennsylvania–New Jersey–Maryland (PJM). Together, these datasets cover electricity prices, electricity demand, bicycle rental demand, and retail sales. We use chronological data splits and restrict inputs to information available at each forecast origin. A historical case is eligible for retrieval only after its complete forecast horizon has been observed. The original and SCoRe-enhanced versions of each backbone are evaluated on the same test windows and forecasting horizons.

Baselines. We evaluate SCoRe with eight forecasting backbones, ranging from the linear model DLinear (Zeng et al., 2023) to Transformer-based models TimeXer (Wang et al., 2024), iTransformer (Liu et al., 2024), PatchTST (Nie et al., 2022), and Temporal Fusion Transformer (Lim et al., 2021) (TFT), followed by retrieval-augmented forecasting methods TS-RAG (Ning et al., 2026) and RAFT (Han et al., 2025), and the pretrained forecasting model Chronos-2 (Ansari et al., 2025). For each backbone, we compare its original predictions with those obtained by adding SCoRe. The base forecaster remains frozen during SCoRe training, and both versions use the same base-model checkpoint and input configuration. This paired comparison isolates the effect of the proposed plugin.

Metrics. We evaluate forecasting performance using mean squared error (MSE) and mean absolute error (MAE), with lower values indicating better performance (Hyndman & Koehler, 2006). Each backbone and its SCoRe-enhanced counterpart use identical target scaling. We report the relative error reduction over the backbone as Gain, with positive values indicating improvement and negative values indicating degradation (Appendix A.4).

Table 1: Forecasting performance of eight backbones before and after applying SCoRe on the five EPF subsets, EDF, and Bike Sharing. Lower MSE and MAE are better. Gain denotes the relative error reduction over the original backbone; positive values indicate improvement.

Backbone	Setting	EPF										EDF		Bike Sharing	
		BE		DE		FR		NP		PJM					
		MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓
DLinear	Original	0.440	0.313	0.508	0.458	0.402	0.240	0.298	0.316	0.091	0.197	0.132	0.257	0.434	0.445
	+SCoRe	0.422	0.295	0.471	0.446	0.394	0.226	0.283	0.311	0.085	0.187	0.106	0.237	0.420	0.434
	Gain (%)	4.09	5.75	7.28	2.62	1.99	5.83	5.03	1.58	6.59	5.08	19.70	7.78	3.23	2.47
TimeXer	Original	0.394	0.266	0.330	0.360	0.381	0.210	0.256	0.286	0.101	0.204	0.104	0.241	0.335	0.388
	+SCoRe	0.380	0.260	0.315	0.348	0.348	0.194	0.224	0.267	0.079	0.183	0.094	0.229	0.325	0.378
	Gain (%)	3.55	2.26	4.55	3.33	8.66	7.62	12.50	6.64	21.78	10.29	9.62	4.98	2.99	2.58
iTransformer	Original	0.388	0.264	0.489	0.457	0.425	0.216	0.313	0.337	0.089	0.188	0.120	0.255	0.543	0.527
	+SCoRe	0.381	0.259	0.465	0.437	0.401	0.212	0.300	0.327	0.083	0.181	0.098	0.233	0.411	0.443
	Gain (%)	1.80	1.89	4.91	4.38	5.65	1.85	4.15	2.97	6.74	3.72	18.33	8.63	24.31	15.94
PatchTST	Original	0.434	0.281	0.543	0.475	0.479	0.239	0.333	0.336	0.151	0.265	0.125	0.254	0.359	0.399
	+SCoRe	0.404	0.268	0.490	0.452	0.398	0.220	0.312	0.324	0.117	0.222	0.106	0.237	0.332	0.377
	Gain (%)	6.91	4.63	9.76	4.84	16.91	7.95	6.31	3.57	22.52	16.23	15.20	6.69	7.52	5.51
TFT	Original	0.411	0.273	0.493	0.443	0.437	0.227	0.306	0.304	0.108	0.197	0.119	0.252	0.323	0.377
	+SCoRe	0.381	0.265	0.432	0.417	0.367	0.205	0.249	0.276	0.096	0.182	0.107	0.237	0.260	0.334
	Gain (%)	7.30	2.93	12.37	5.87	16.02	9.69	18.63	9.21	11.11	7.61	10.08	5.95	19.50	11.41
TS-RAG	Original	0.340	0.228	0.437	0.408	0.365	0.185	0.227	0.252	0.075	0.170	0.082	0.199	0.312	0.321
	+SCoRe	0.338	0.227	0.415	0.401	0.365	0.184	0.227	0.255	0.075	0.169	0.072	0.191	0.308	0.319
	Gain (%)	0.59	0.44	5.03	1.72	0.00	0.54	0.00	-1.19	0.00	0.59	12.20	4.02	1.28	0.62
RAFT	Original	0.501	0.332	0.502	0.455	0.488	0.272	0.300	0.314	0.107	0.214	0.173	0.306	0.557	0.535
	+SCoRe	0.462	0.298	0.456	0.427	0.459	0.246	0.277	0.300	0.096	0.200	0.151	0.288	0.462	0.486
	Gain (%)	7.78	10.24	9.16	6.15	5.94	9.56	7.67	4.46	10.28	6.54	12.72	5.88	17.06	9.16
Chronos-2	Original	0.285	0.212	0.237	0.315	0.294	0.158	0.183	0.217	0.068	0.159	0.142	0.291	0.348	0.363
	+SCoRe	0.282	0.206	0.236	0.305	0.292	0.159	0.182	0.217	0.067	0.158	0.135	0.275	0.317	0.345
	Gain (%)	1.05	2.83	0.42	3.17	0.68	-0.63	0.55	0.00	1.47	0.63	4.93	5.50	8.91	4.96

Table 2: Ablation study on EDF and Bike Sharing (BS). Checkmarks indicate enabled modules.

SRM	CRT	DUE	RCB	EDF		BS	
				MSE	MAE	MSE	MAE
				0.104	0.241	0.335	0.388
✓				0.100	0.236	0.332	0.384
✓	✓			0.099	0.233	0.330	0.381
✓	✓	✓		0.096	0.232	0.327	0.380
✓	✓	✓	✓	**0.094**	**0.229**	**0.325**	**0.378**

Table 3: Forecasting performance on M5 with and without SCoRe.

Backbone	MSE			MAE		
	Base	+ SCoRe	Gain (%)	Base	+ SCoRe	Gain (%)
DLinear	3.04	2.88	5.26	1.06	0.98	7.55
PatchTST	2.93	2.85	2.73	1.06	1.01	4.72
TimeXer	2.90	2.63	9.31	1.05	0.90	14.29
Chronos-2	3.61	3.11	13.85	1.06	0.92	13.21

4.2 Main Results

Table 1 shows that SCoRe improves forecasting performance across eight backbones and seven evaluation settings. At the reported precision, it reduces MSE in 53 of the 56 combinations of backbones and datasets and MAE in 53, with average relative reductions of 8.31% and 5.10%, respectively. MSE remains unchanged in the other three combinations, while MAE remains unchanged in one and increases in two. All six backbones other than TS-RAG and Chronos-2 improve on both metrics across every setting. Gains are generally smaller for TS-RAG and Chronos-2, with occasional MAE degradation, indicating that the benefit depends on the backbone and dataset. Results on M5 further demonstrate the effectiveness of SCoRe in retail demand forecasting. **In a real-world industrial retail setting**, adding SCoRe to the production model increases ACC from 67.44% to 71.45%, an absolute improvement of 4.01 percentage points. ACC measures forecasting accuracy and is defined in Appendix A.4. Overall, these results support SCoRe as a broadly applicable forecasting plugin with practical value in retail applications.

4.3 Ablation Studies

Table 2 presents a progressive ablation of SCoRe on EDF and BS. Introducing SRM improves both MSE and MAE over the backbone, while subsequently adding CRT and DUE further reduces

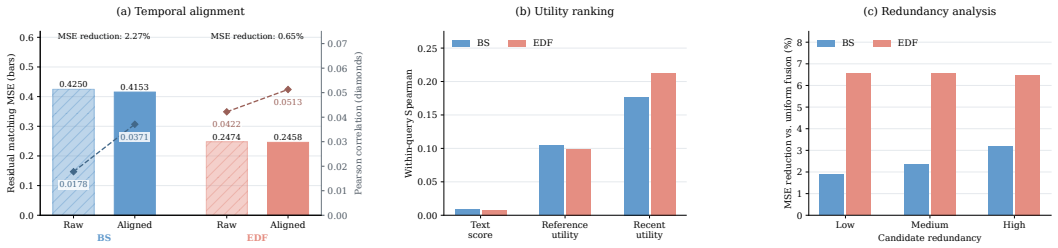


Figure 3: Mechanism analysis with TimeXer on BS and EDF. (a) Residual matching before and after alignment. (b) Rank correlation between candidate scores and realized directional gains within each query. (c) MSE reduction over adapted uniform fusion across redundancy levels.

forecasting errors on both datasets. Finally, enabling RCB lowers MSE from 0.096 to 0.094 on EDF and from 0.327 to 0.325 on BS, yielding the best performance across all four metrics. The full model reduces MSE by 9.62% on EDF and 2.99% on BS relative to the backbone. These incremental gains support the complementary roles of all four components in this configuration.

4.4 Analysis Studies

We analyze temporal alignment, utility ranking, and candidate fusion using TimeXer on BS and EDF, averaging results over three random seeds. Test targets are used only for evaluation.

Temporal alignment. We compare raw and aligned residuals against $\mathbf{e}_q$ using matching MSE and Pearson correlation, averaged first within each query and then across queries. Figure 3(a) shows that alignment reduces matching MSE by 2.27% on BS and 0.65% on EDF, while increasing correlation from 0.02 to 0.04 and from 0.04 to 0.05, respectively. These results indicate improved residual compatibility, rather than directly measuring forecasting gains.

Utility ranking. For the same transported candidates $\mathbf{v}_{qj}$, we measure Spearman correlation between candidate scores and g^*_{qj} (Eq. (4)) within each query. Figure 3(b) shows that correlations on BS and EDF rise from 0.01 and 0.01 for text scores to 0.10 and 0.10 for reference utility, and to 0.18 and 0.21 for recently adapted utility. DUE thus improves ranking beyond semantic relevance, although forecasting gains also depend on correction magnitude and candidate interactions.

Candidate redundancy. We group test queries by the average positive pairwise correlation among transported residuals, using tertile thresholds derived from training data. Compared with uniform fusion of the same aligned and adapted candidates, SCoRe reduces MSE by 1.88%, 2.34%, and 3.17% on BS as redundancy increases, and by 6.55%, 6.58%, and 6.48% on EDF (Figure 3(c)). These results support the full fusion procedure across redundancy levels, particularly at higher redundancy on BS, but do not isolate the effects of Gram interactions or RCB.

4.5 Case Studies

Case 1: Semantic retrieval. Using the same memory and temporal constraints, semantic retrieval finds a reference whose zero sales proportion is closer to the query (46.4% vs 48.2%) than numerical

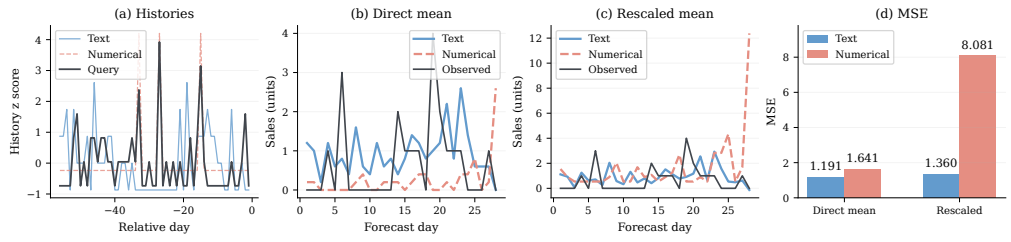


Figure 4: **Case 1: Semantic versus numerical retrieval on M5.** (a) Standardized histories of the query and each method's top reference. (b–c) Mean continuations of the top five references, before and after historical rescaling, compared with observed sales. (d) Corresponding MSE, evaluating retrieval quality without learned adaptation.

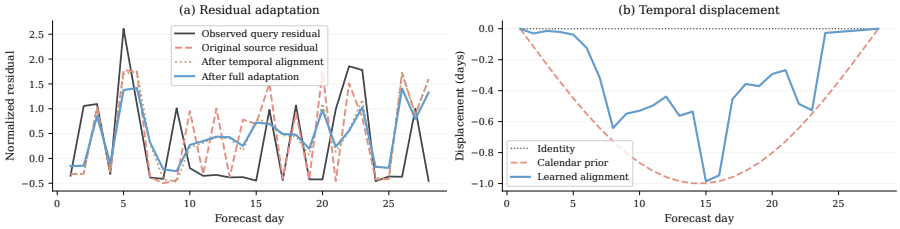


Figure 5: **Case 2: Response transport on M5.** (a) A fixed source residual before alignment, after temporal alignment, and after full adaptation, compared with the observed query residual. (b) Temporal displacement under the calendar prior and learned alignment, with endpoints fixed.

retrieval (94.6%), indicating a more compatible historical demand state. Averaging the five highest ranked continuations reduces MSE from 1.64 to 1.19 (27.4%; Figure 4). This measures retrieval quality, not complete model performance (Appendix G.1).

Case 2: Response transport. For the same reference, temporal alignment reduces normalized residual MSE from 0.98 to 0.87; magnitude and shape adaptation further reduce it to 0.79 (20.1% overall; Figure 5). The learned mapping adjusts local timing while preserving the endpoints, with a maximum displacement of 0.99 days. This is a reconstruction gain, not an equivalent forecast improvement. Details are provided in Appendix G.2.

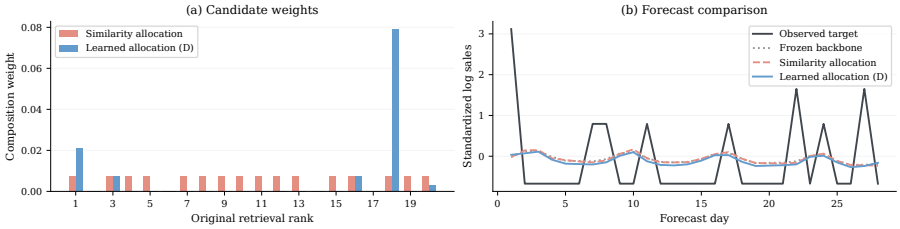


Figure 6: **Case 3: Utility composition on M5.** (a) Weights assigned to fixed adapted candidates by similarity weighting and utility composition in branch D. (b) Forecasts under both weighting schemes, compared with the frozen backbone and observed target.

Case 3: Utility composition. Figure 6 compares similarity weighting with utility composition while keeping the adapted candidates, backbone, query gate, and total weight fixed. Similar cosine scores produce nearly uniform weights, whereas branch D concentrates weight on fewer references, favoring retrieval rank 18. Utility composition achieves a forecast MSE of 0.96, compared with 0.99 for similarity weighting, a reduction of 2.68%. See Appendix G.3 for details.

5 Conclusion

In this paper, we presented SCoRe, a plug-and-play framework that connects semantic retrieval with conditional residual transfer for time series forecasting. Joint-context semantic descriptions identify potentially transferable historical cases beyond direct numerical similarity. Temporal alignment and low-rank adaptation transform their forecast residuals into target-specific corrections, while utility-based joint selection assesses their combined predictive value and manages negative transfer risk. Experiments across diverse forecasting backbones and datasets demonstrate the broad effectiveness of this approach. Our findings highlight that semantic relevance is a starting point for transfer, whose predictive value depends on explicit numerical adaptation and selection. SCoRe offers a practical way to connect these stages and use historical forecasting experience to improve existing models. Nevertheless, SCoRe depends on the quality of LLM-generated descriptions and the availability of informative historical cases, which may limit its effectiveness under poorly represented contexts. Its utility estimates and validation-based safeguards also do not guarantee improvements under unseen distribution shifts. Future work will investigate uncertainty-aware retrieval and correction selection, together with adaptive memory updates, to improve robustness as forecasting conditions evolve.

AI USE STATEMENT

We used generative AI tools to support research ideation and execution, draft sections of the manuscript, and improve the clarity and readability of the writing. We also used these tools to assist in developing and checking proofs of mathematical claims.

The authors reviewed and revised all AI-assisted content, assessed methodological suggestions against the experimental evidence, and checked mathematical arguments for correctness and consistency with their assumptions. All reported experimental results were obtained from the actual experiments. The authors take full responsibility for the final manuscript, including all text, methods, mathematical claims, and results produced with the assistance of generative AI.

ETHICS STATEMENT

Our experiments use publicly available benchmark datasets and an internal dataset from a real-world e-commerce platform. The proposed forecasting framework is intended to support applications such as demand forecasting and inventory planning. No human subjects were involved in this research.

REPRODUCIBILITY STATEMENT

The performance of SCoRe and datasets used in our work are real, and all experimental results can be reproduced. The code for SCoRe is available.

REFERENCES

Ernesto Aguilar Madrid and Nuno Antonio. Short-term electricity load forecasting with machine learning. *Information*, 12(2):50, 2021.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Vitor Cerqueira, Luis Torgo, and Igor Mozetič. Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, 109(11):1997–2028, 2020.

Mouxiang Chen, Lefei Shen, Han Fu, Zhuo Li, Jianling Sun, and Chenghao Liu. Calibration of time-series forecasting: Detecting and adapting context-driven distribution shift. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 341–352, 2024.

Afrin Dange and Sunita Sarawagi. Tfmadapter: Lightweight instance-level adaptation of foundation models for forecasting with covariates. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 498–507, 2025.

Abhimanyu Das, Weihao Kong, Andrew Leach, Shaan Mathur, Rajat Sen, and Rose Yu. Long-term forecasting with tide: Time-series dense encoder. *arXiv preprint arXiv:2304.08424*, 2023.

Yuntao Du, Jindong Wang, Wenjie Feng, Sinno Pan, Tao Qin, Renjun Xu, and Chongjun Wang. Adarnn: Adaptive learning and forecasting of time series. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pp. 402–411, 2021.

H Fanaee-T. Bike sharing-uci machine learning repository, 2013.

Robert Fildes, Shaohui Ma, and Stephan Kolassa. Retail forecasting: Research and practice. *International Journal of Forecasting*, 38(4):1283–1318, 2022.

Lu Han, Yu Liu, Lan Li, Qiwen Deng, Jian Jiang, Zhe Yu, Binfeng Wang, Xingyu Lu, Lintao Ma, Han-Jia Ye, et al. Unica: Unified covariate adaptation for time series foundation model. In *International Conference on Learning Representations*, volume 2026, pp. 6403–6443, 2026.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan O Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. *arXiv preprint arXiv:2505.04163*, 2025.

Yuchen He, Yueyang Cang, Zhiyuan Ning, Ningyu Wang, and Li Shi. Ratl: Learning from retrieved residuals for robust multivariate time-series forecasting. *arXiv preprint arXiv:2609.03937*, 2026.

Tao Hong and Shu Fan. Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, 32(3):914–938, 2016.

Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 427–435. PMLR, 2013. URL https://proceedings.mlr.press/v28/jaggi13.html.

HyunGi Kim, Siwon Kim, Jisoo Mok, and Sungroh Yoon. Battling the non-stationarity in time series forecasting via test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17868–17876, 2025.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*, 7, 1994.

Jesus Lago, Grzegorz Marcjasz, Bart De Schutter, and Rafał Weron. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, 293:116983, 2021.

Kenny Ye Liang, Zhongyi Pei, Huan Zhang, Yuhui Liu, Shaoxu Song, and Jianmin Wang. Retrieval-augmented generation with covariate time series. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 7624–7635, 2026.

Bryan Lim, Sercan Ö 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.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 11116–11140, 2024.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m5 competition: Background, organization, and implementation. *International Journal of Forecasting*, 38(4):1325–1336, 2022.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Kanghui Ning, Zijie Pan, Yu Liu, Yushan Jiang, James Zhang, Kashif Rasul, Anderson Schneider, Lintao Ma, Yuriy Nevmyvaka, and Dongjin Song. Ts-rag: Retrieval-augmented generation based time series foundation models are stronger zero-shot forecaster. *Advances in Neural Information Processing Systems*, 38:163170–163199, 2026.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

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.

Yihong Tang, Tong Nie, Junlin He, Qianjun Huang, Dingyi Zhuang, and Lijun Sun. Bridge: Retrieval-augmented spatiotemporal modeling for urban delivery demand. *arXiv preprint arXiv:2605.19172*, 2026.

Eleni I Vlahogianni, Matthew G Karlaftis, and John C Golias. Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C: Emerging Technologies*, 43: 3–19, 2014.

Siyuan Wang, Peng Chen, Yihang Wang, Wanghui Qiu, Guo Guo, Bin Yang, and Yang Shu. Unlocking the value of text: Event-driven reasoning and multi-level alignment for time series forecasting. In *International Conference on Learning Representations*, volume 2026, pp. 31182–31210, 2026.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. *Advances in neural information processing systems*, 37:469–498, 2024.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11285–11294. IEEE, 2019.

Qingsong Wen, Weiqi Chen, Liang Sun, Zhang Zhang, Liang Wang, Rong Jin, Tieniu Tan, et al. Onenet: Enhancing time series forecasting models under concept drift by online ensembling. *Advances in Neural Information Processing Systems*, 36:69949–69980, 2023.

Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados, et al. Context is key: A benchmark for forecasting with essential textual information. *arXiv preprint arXiv:2410.18959*, 2024.

David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. *arXiv preprint arXiv:2402.02592*, 2024.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024. URL https://arxiv.org/abs/2309.07597.

Hao Xue and Flora D Salim. Promptcast: A new prompt-based learning paradigm for time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6851–6864, 2023.

Yosuke Yamaguchi, Issei Suemitsu, and Wenpeng Wei. Citras: Covariate-informed transformer for time series forecasting. *IEEE Access*, 2026.

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.

Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 10(2):305–329, 2022.