TS-RAG: Retrieval-Augmented Generation based Time Series Foundation Models are stronger Zero-Shot Forecaster

Kanghui Ning[1], Zijie Pan[1], Yu Liu[3], Yushan Jiang[1]
James Y.Zhang[3], Kashif Rasul[2], Anderson Schneider[2]
Lintao Ma[3], Yuriy Nevmyvaka[2], Dongjin Song[1]
[1]School of Computing, University of Connecticut
[2]Department of Machine Learning Research, Morgan Stanley
[3]Ant Group

Introduction

- Background: Time Series Foundation Models (**TSFMs**) achieve strong **zero-shot forecasting**.
- Target: Propose a retrieval-augmented framework to improve **generalization** and **interpretability** of TSFMs for zero-shot forecasting.
- Approach: A **lightweight**, **plug-and-play** module that enhances TSFMs via **post-training**, with a **customizable knowledge base**.

Motivation

Limitation of existing models:

- Generalization: LLMs and TSFMs still face challenge in domain generalization and distribution shifts.
- Interpretability: TSFMs also suffer from poor interpretability for explaining their predictions.

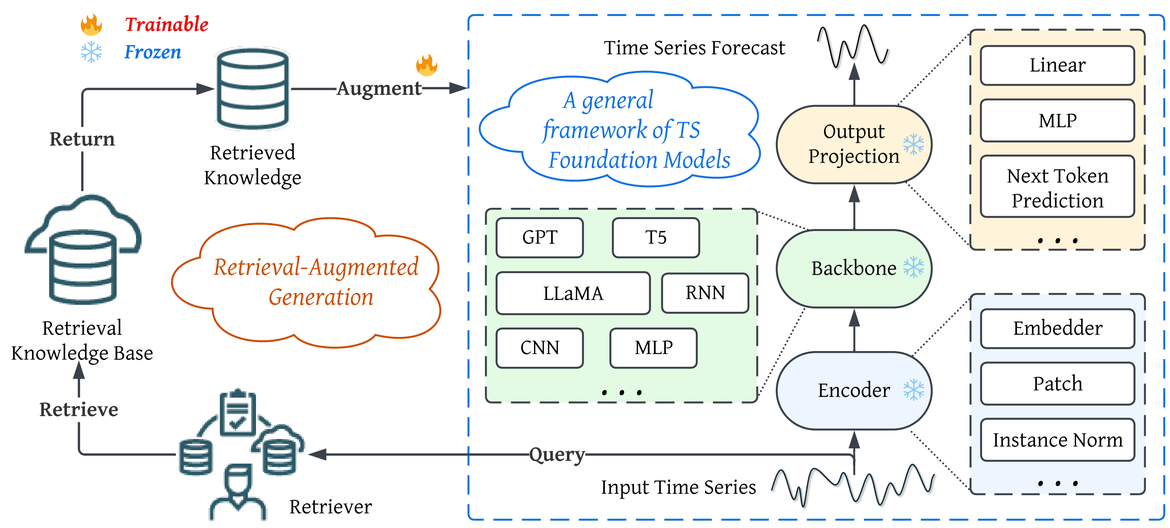
Opportunity in Retrieval-Augmented Generation:

- RAG in NLP: proven to boost adaptability and accuracy.
- Time series data often exhibit similar patterns across history, domains or datasets.

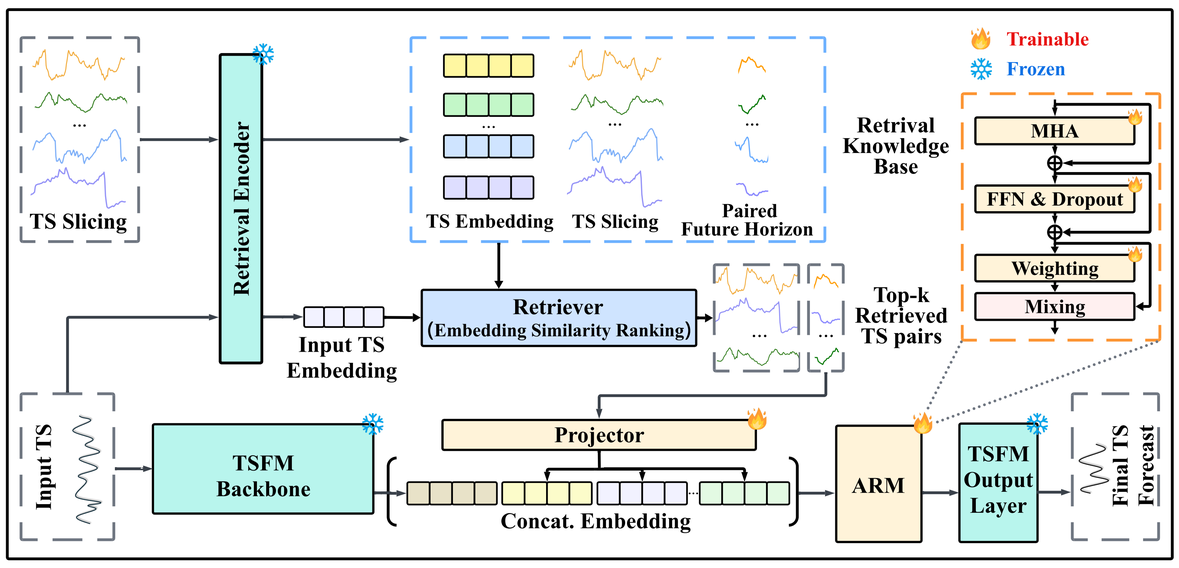
Contribution

- We propose **TS-RAG**, a RAG based time series forecasting **framework** for **TSFMs**.
- We develop a learnable **Adaptive Retrieval Mixer (ARM)** augmentation module that can dynamically enhance the time series representation.
- Experiments on seven real-world datasets show that TS-RAG improves performance by **up to 6.8%** over state-of-the-art TSFMs while also providing **interpretability** for its predictions.

Methodology



Overall Framework of TS-RAG



Detailed Architecture of the Proposed Framework

TS-RAG consists of three key components: a **Time Series Foundation Model** (TSFM), a **Retriever** and an **Adaptive Retrieval Mixer (ARM)** augmentation module.

- The Retriever calculates the Euclidean distance between the query embedding and each stored context embedding in the knowledge base, and then selects the top-k similar candidates based on the smallest distance.

$$\mathbf{e}_q = f_{\text{enc}}(\mathbf{x}_q). \quad (1) \qquad d(\mathbf{e}_q, \mathbf{e}_i) = \|\mathbf{e}_q - \mathbf{e}_i\|_2, \quad \forall i \in \{1, 2, \ldots, n\}. \quad (2)$$

$$\mathcal{C} = \text{TopK}_{\text{min}}\left(\{(\mathbf{x}_i, \mathbf{y}_i, d(\mathbf{e}_q, \mathbf{e}_i)) \mid i = 1, 2, \ldots, n\}, k\right), \quad (3)$$

- To perform forecasting, The ARM dynamically weight and fuse each embedding of the top-k retrieved forecasting horizons.

$$\hat{\mathbf{e}}_i = f_{\text{MLP}}(\mathbf{y}_i), \quad i = 1, 2, \ldots, k \quad (4) \qquad E_{\text{ret}} = [\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \ldots, \hat{\mathbf{e}}_k] \in \mathbb{R}^{k \times d}, \quad (5)$$

$$E_{\text{concat}} = [\hat{\mathbf{e}}_q; E_{\text{ret}}] \in \mathbb{R}^{(k+1) \times d}. \quad (6) \qquad E_{\text{att}} = \text{MHA}(E_{\text{concat}}) + E_{\text{concat}}, \quad (7)$$

$$E_{\text{ffn}} = \text{Dropout}(\text{FFN}(E_{\text{att}})) + E_{\text{att}}. \quad (8) \qquad \alpha = \text{Softmax}(W_g E_{\text{ffn}} + b_g). \quad (9)$$

$$\mathbf{e}_{\text{final}} = \hat{\mathbf{e}}_q + \sum_{i=1}^{k+1} \alpha_i E_{\text{ffn},i}. \quad (10) \qquad \hat{\mathbf{y}}_q = f_{\text{proj}}(\mathbf{e}_{\text{final}}), \quad (11)$$

Results

Methods	TS-RAG$_{\text{Chronos-bolt}}$		Chronos-bolt$_{\text{B}}$		MOMENT		TTM$_{\text{B}}$		Moirai$_{\text{B}}$		TimesFM		Chronos$_{\text{B}}$	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	**0.3557**	**0.3624**	0.3616	0.3650	0.3920	0.4110	0.3619	0.3710	0.3686	0.3835	0.4254	0.3825	0.4217	0.3806
ETTh2	**0.2451**	**0.2982**	0.2517	0.2992	0.2742	0.3327	0.2531	0.3032	0.2547	0.3053	0.2894	0.3233	0.2659	0.3136
ETTm1	**0.2906**	**0.3114**	0.3109	0.3185	0.3506	0.3834	0.3152	0.3248	0.5399	0.4322	0.3321	0.3326	0.3935	0.3695
ETTm2	**0.1466**	**0.2231**	0.1487	0.2236	0.1703	0.2579	0.1511	0.2405	0.1958	0.2687	0.1703	0.2552	0.1663	0.2522
Weather	**0.1454**	**0.1771**	0.1525	0.1825	0.1801	0.2384	0.1543	0.1893	0.1711	0.1912	—	—	0.1897	0.2107
Electricity	**0.1120**	**0.2002**	0.1132	0.2004	0.1967	0.3028	0.1715	0.2643	0.1832	0.2814	—	—	0.1460	0.2237
Exchange rate	**0.0627**	**0.1718**	0.0673	0.1780	0.0979	0.2059	0.0657	0.1725	0.0663	0.1720	0.0695	0.1802	0.0831	0.1879

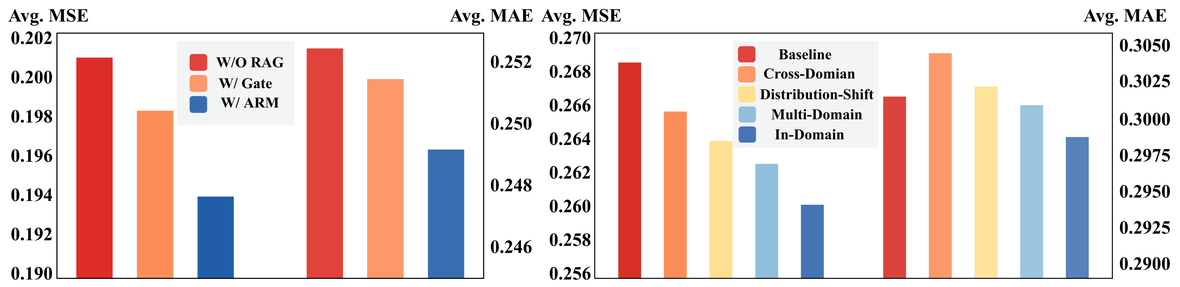
Long-term Zero-shot Forecasting Results

Method	Metric	ETTh1	ETTh2	ETTm1	ETTm2	Weather	Electricity	Exchange	Average
TS-RAG$_{\text{ARM}}$	MSE	**0.3557**	**0.2451**	**0.2906**	**0.1466**	**0.1454**	**0.1120**	**0.0627**	**0.1940**
	MAE	**0.3624**	**0.2982**	**0.3114**	**0.2231**	**0.1771**	**0.2002**	**0.1718**	**0.2492**
TS-RAG$_{\text{Gate}}$	MSE	0.3575	0.2498	0.3041	0.1473	0.1501	0.1126	0.0663	0.1982
	MAE	0.3640	0.2988	0.3154	0.2235	0.1815	0.2005	0.1768	0.2515
Chronos-bolt	MSE	0.3616	0.2517	0.3109	0.1487	0.1525	0.1132	0.0673	0.2008
	MAE	0.3650	0.2992	0.3185	0.2236	0.1825	0.2004	0.1780	0.2525

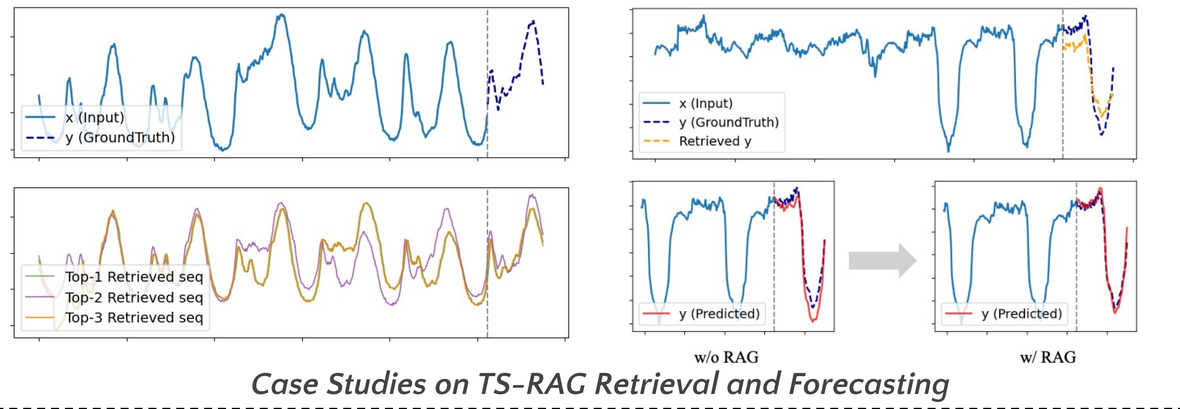
Ablation Study on Augmentation Modules

Backbone	Metric	ETTh1	ETTh2	ETTm1	ETTm2	Weather	Electricity	Exchange	Average
Chronos-bolt	MSE	0.3616	0.2517	0.3109	0.1487	0.1525	0.1132	0.0673	0.2008
	MAE	0.3650	0.2992	0.3185	0.2236	0.1825	0.2004	0.1780	0.2525
TS-RAG$_{\text{Chronos-bolt}}$	MSE	**0.3557**	**0.2451**	**0.2906**	**0.1466**	**0.1454**	**0.1120**	**0.0627**	**0.1940**
	MAE	**0.3624**	**0.2982**	**0.3114**	**0.2231**	**0.1771**	**0.2002**	**0.1718**	**0.2492**
Moment	MSE	0.3920	0.2742	0.3506	0.1703	0.1801	0.1967	0.0979	0.2374
	MAE	0.4110	0.3327	0.3824	0.2579	0.2384	0.3028	0.2059	0.3044
TS-RAG$_{\text{Moment}}$	MSE	**0.3823**	**0.2511**	**0.3325**	**0.1552**	**0.1604**	**0.1920**	**0.0775**	**0.2216**
	MAE	**0.4072**	**0.3220**	**0.3738**	**0.2474**	**0.2212**	**0.2994**	**0.1972**	**0.2955**

Zero-shot Forecasting with TS-RAG across Different Backbones



Ablation Studies on Augmentation Methods (left) and Knowledge Bases (right)



Case Studies on TS-RAG Retrieval and Forecasting

TS-RAG improves forecasting interpretability in two ways:

- *Retrieval-as-evidence*, which surfaces the top-k relevant sequences for each query window.
- *Transparent weighting*, which highlights the most influential retrieved sequences based on similarity scores.