Align-RAG: Alignment Is (Almost) All You Need for TSFM In-Context Learning

Mohammad Asadi[1] **Soheil Hor**[2] **Bardiya Akhbari**[2]
Jack W. O'Sullivan[1] **Tahoura Nedaee**[1] **Layne C. Price**[2]
Raviteja Anantha[2] **Euan Ashley**[1] **Ehsan Adeli**[1]

[1]Stanford University [2]Amazon

masadi@stanford.edu

Abstract

Retrieval-augmented forecasting promises to adapt frozen Time Series Foundation Models (TSFMs) to new domains without fine-tuning, but recent methods typically rely on learned fusion modules, i.e., trained adapters that merge retrieved examples into the backbone's forecast, based on the assumption that frozen backbones cannot dynamically incorporate retrieved context on their own. We show this assumption is unnecessary. We introduce **Align-RAG**, a training-free method that applies a closed-form per-pair amplitude rescaling and integer-lag phase shift to retrieved past–future windows before they enter a frozen backbone's context. With no learned parameters, Align-RAG outperforms the state-of-the-art trained retrieval adapter on a frozen Chronos-Bolt on all seven datasets of the standard benchmark (avg -3.75% MSE), showing that the gains previously attributed to learned fusion are recoverable without any training. Align-RAG further improves zero-shot MSE on four additional frozen TSFMs with various architectures by 2.5% to 13.7% per backbone with no per-backbone tuning. To probe why alignment helps, we compare the frozen backbone's prediction shift under aligned demonstrations to the closed-form ridge prediction shift on the same pairs. We find that aligned demonstrations induce prediction shifts that track a closed-form ridge predictor on the same pairs, with a future-shuffle control ruling out a futures-averaging account. Together, these results indicate that frozen TSFMs already support dynamic in-context use of retrievals, and that closed-form alignment should be the default baseline for retrieval-augmented forecasting before any fusion module is trained. Code available at: `https://github.com/masadi-99/align-rag`

1 Introduction

Time series forecasting underpins decision-making in domains as varied as energy management, finance, healthcare, and climate science. Time Series Foundation Models (TSFMs) such as Chronos [Ansari et al., 2024], Chronos-2 [Ansari et al., 2025], TimesFM [Das et al., 2024], Moirai [Woo et al., 2024], and Toto [Cohen et al., 2024] have established a zero-shot forecasting paradigm by learning shared temporal representations across heterogeneous corpora. Once pretrained, a frozen TSFM can be applied to unseen series without further training, but its accuracy on a target distribution depends on whatever the pretraining corpus exposed.

Retrieval-augmented generation [Lewis et al., 2020] addresses an analogous gap for large language models by providing relevant context at inference time. The corresponding move for time series is to retrieve past–future windows from a reference set whose dynamics resemble the query and supply them to the model as additional context. For example, a hospital deploying a frozen TSFM

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

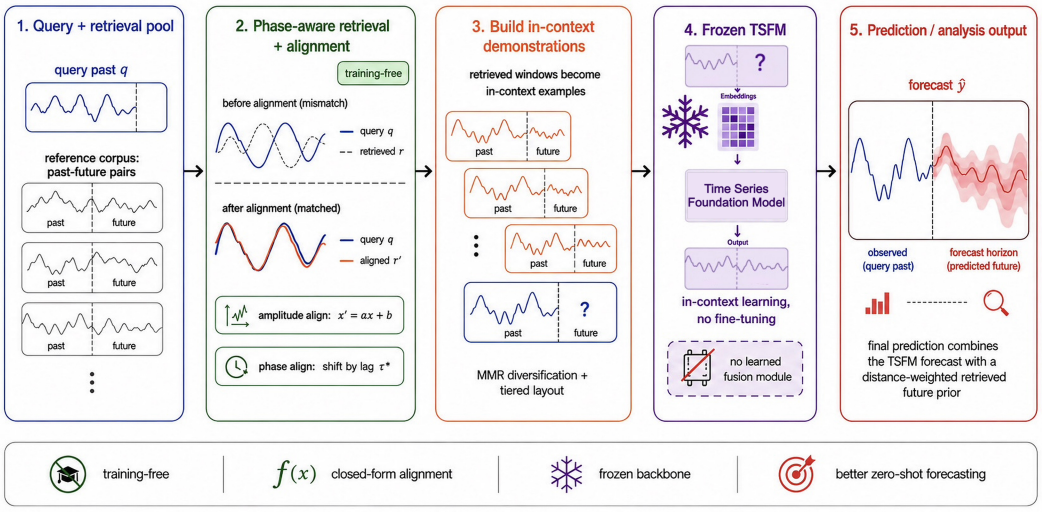


Figure 1: Align-RAG overview. (1) For each query past $\mathbf{q}$, retrieve top-K past–future windows from a reference corpus. (2) Apply two closed-form, training-free transforms per pair so that each retrieved past matches the query while the past–future correspondence is preserved. (3) Pack the aligned pairs as in-context demonstrations using MMR diversification and a tiered token-budget layout. (4) A single forward pass through a frozen Time Series Foundation Model. No learned fusion module is introduced. (5) The final forecast combines the model's in-context output with a distance-weighted prior built from the retrieved aligned futures.

might retrieve patterns from related ICU traces and an energy operator might retrieve consumption windows from similar load regimes. Done well, this offers dynamic adaptation without labelled data or per-domain fine-tuning.

Recent retrieval-augmented forecasting on frozen TSFMs has converged on a single design pattern: pair the backbone with a learned fusion module that mediates between retrievals and the model's representation. TS-RAG (NeurIPS 2025) [Ning et al., 2025], the state-of-the-art in retrieval-augmented forecasting for time series foundation models to the best of our knowledge, states the position cleanly, claiming that pretrained TSFMs "lack inherent mechanisms for domain adaptation, as they cannot incorporate external contextual knowledge dynamically", and trains an Adaptive Retrieval Mixer to fuse encoded retrievals with the backbone. TimeRAF [Zhang et al., 2025] trains an end-to-end retriever and a Channel Prompting module; RATD [Liu et al., 2024] trains a reference-modulated attention block inside a diffusion denoiser; RAFT [Han et al., 2025] trains projection and prediction heads on top of the retrieved patterns; TimesFM-ICF [Faw et al., 2025] re-pretrains the foundation model itself to consume in-context retrieval examples. The underlying assumption is uniform: a learned ad-hoc element is required to make a frozen TSFM benefit from retrieved context. We challenge this assumption.

We introduce Align-RAG, a training-free retrieval-augmented forecasting method that extends a frozen TSFM by aligning retrieved past–future windows to the query in amplitude and phase before they enter the backbone's context. Given a query window, Align-RAG retrieves nearest neighbours from a reference set of (past, future) pairs and applies a closed-form per-pair amplitude rescaling and phase shift so that each retrieved past matches the query. The aligned pairs are then provided as in-context demonstrations to the backbone, whose parameters remain unchanged.

Align-RAG outperforms TS-RAG on a frozen Chronos-Bolt backbone on $7/7$ datasets in MSE and $6/7$ in MAE across the TS-RAG benchmark protocol (ETTh1, ETTh2, ETTm1, ETTm2, Weather, Exchange, Electricity; input length 512, horizon 64), with an average MSE reduction of 3.75% and gains significant under a paired moving-block bootstrap on 5 of 6 small-dataset comparisons. The same procedure transfers across four additional frozen TSFMs (Chronos-2, TimesFM-2.0, Moirai, Toto), improving zero-shot MSE by between 2.5% and 13.7% per backbone, with the largest gains on Moirai (avg -13.7%, single cells reaching -30% on ETTm1 and -20% on Weather), and without any per-backbone tuning. The comparison is non-trivial: TS-RAG's Adaptive Retrieval Mixer is

trained on 26 million context–horizon pairs sampled from Chronos's pretraining corpus [Ning et al., 2025], while Align-RAG introduces no learned parameters at all.

A 2×2 ablation isolates where the gain comes from. Holding the backbone and benchmark fixed, we cross retriever (trained TS-RAG retriever vs. random retrieval) with alignment (on vs. off). Switching from the trained retriever to a random retriever, with alignment on, costs 0.84 percentage points of MSE on average and still beats the trained TS-RAG mixer on 4 of 7 datasets. Turning alignment off, with either retriever, costs more than 20 percentage points of MSE. The ranking of retrieved windows accounts for under one point of the headline; the alignment of the windows that are retrieved accounts for the rest.

We also probe the mechanism behind this gain. For each test window we compute the cosine similarity ρ_{GD} between the frozen backbone's prediction shift, induced by replacing zero-shot context with aligned demonstrations, and the closed-form ridge prediction shift on the same demonstration pairs. On Chronos-Bolt, ρ_{GD} rises from 0.30 without alignment to 0.45 with alignment, and a paired moving-block bootstrap excludes zero on $7/7$ datasets. As a causal control, we shuffle the futures of the retrieved pairs while leaving the pasts and the alignment intact. On ETTm1, the 4.2% MSE reduction from alignment flips to a 12.2% MSE increase under the future shuffle, in the direction the closed-form ridge predicts on the same shuffled pairs, ruling out an explanation in which the backbone is averaging over retrieved futures and ignoring the past–future relationship.

Contributions.

- Align-RAG, a training-free retrieval-augmented forecasting method that aligns retrieved past–future windows to the query in amplitude and phase before they enter a frozen TSFM's context, without modifying the backbone or training any new component.
- Evaluation against the trained TS-RAG mixer on the standard seven-dataset benchmark with a frozen Chronos-Bolt (Section 4.2) and across four further frozen TSFMs (Section 4.4).
- A behavioural mechanism analysis (Section 5): under our closed-form alignment, the frozen backbone's prediction shift tracks the closed-form ridge prediction shift on the same demonstrations, and a future-shuffle causal control confirms the model is regressing on the (past, future) pairs rather than averaging over their futures.

2 Related work

Retrieval-augmented forecasting for TSFMs. A growing body of work attaches trained components to a frozen TSFM to consume retrievals. TS-RAG [Ning et al., 2025] trains an Adaptive Retrieval Mixer on top of frozen Chronos-Bolt [Ansari et al., 2024]; TimeRAF [Zhang et al., 2025] trains an end-to-end retriever and a Channel Prompting module; RAFT [Han et al., 2025] trains projection and prediction heads; RATD [Liu et al., 2024] trains a diffusion denoiser with a learned reference-modulated attention block; TimeRAG [Yang et al., 2025] inserts a learned reprogramming layer; Advanced RAF [Tire et al., 2024] fine-tunes the backbone end-to-end (while Naive RAF, although a training-free variant, is outperformed substantially by TS-RAG[Ning et al., 2025]). Each of these methods relies on a learned ad-hoc element interposed between the retrieval and the backbone. We compare numerically against TS-RAG, which is the most directly comparable and the strongest trained-fusion baseline at the time of writing; we adopt its protocol throughout for fair comparison. The corresponding move on the LLM side has typically been simpler. Methods such as REPLUG [Shi et al., 2024], kNN-LM [Khandelwal et al., 2020], and In-Context RALM [Ram et al., 2023] concatenate or interpolate retrieved context with a frozen LM directly.

In-context learning as implicit regression. A line of theoretical work frames in-context learning as implicit regression. Garg et al. [2022] show that transformers trained on linear functions match the least-squares estimator; Akyürek et al. [2023] and von Oswald et al. [2023] demonstrate that the algorithm such transformers implement is consistent with gradient descent or closed-form ridge regression on the in-context demonstrations. Faw et al. [2025] extends a related lens to TSFMs by re-pretraining the foundation model to consume in-context retrieval examples. We use this body of work as a behavioural reference: in Section 5 we compare the frozen TSFM's prediction shift under our alignment to the closed-form ridge predictor's shift on the same demonstrations.

Phase, amplitude, and structural similarity in time series. Dynamic time warping [Cuturi and Blondel, 2017] and its shape variants [Zhao and Itti, 2018] compare misaligned series; k-Shape [Paparrizos and Gravano, 2015] clusters by phase-invariant Pearson similarity; the matrix profile [Yeh et al., 2016] indexes motifs under z-normalised distance. These methods treat amplitude and phase as nuisance directions and remove them before downstream learning.

3 Method: Align-RAG

Align-RAG operates on each retrieved past–future window, before tokenization, with two closed-form transforms. The first rescales the retrieved past so its first and second moments agree with the query's, through a regularised affine fit. The second shifts the retrieved past in time so its principal lag with respect to the query is zero, as an integer-sample shift chosen by cross-correlation. Both transforms have the same future-side counterparts: the same affine map and the same lag, applied to the retrieved future so the $(\mathbf{p}_i, \mathbf{f}_i)$ pair stays consistent. Neighbour diversification, packing into the backbone's context budget, future blending, and a no-demonstration second forward pass are further training-free components that complement the alignment and improve zero-shot accuracy.

Let $\mathbf{q} \in \mathbb{R}^S$ denote the query past of length S (S=512 for Chronos-Bolt) and $\mathbf{y} \in \mathbb{R}^H$ its target future of horizon H. A retrieval index returns a pool $\mathcal{P} = \{(\mathbf{p}_i, \mathbf{f}_i, d_i)\}_{i=1}^{20}$ of (past, future, distance) triples per query channel. We write $f_\theta(C)$ for the frozen TSFM's forecast given a context C.

3.1 Amplitude alignment

For each retrieved neighbour, indexed by i, we fit a single affine map on the past and apply it to both past and future. Let μ_q, σ_q denote the mean and standard deviation of the query past $\mathbf{q}$, and μ_{p_i}, σ_{p_i} those of the retrieved past $\mathbf{p}_i$. The Wiener-style shrunk slope a_i and intercept b_i are

$$a_i \;=\; \frac{\sigma_q\,\sigma_{p_i}}{\sigma_{p_i}^2 + (\sigma_q/M)^2}, \qquad b_i \;=\; \mu_q - a_i\,\mu_{p_i} \tag{1}$$

where a_i is the regularised counterpart of the plug-in scale ratio σ_q/σ_{p_i}, in the spirit of Wiener filtering for noisy estimation [Wiener, 1949]. The shrinkage parameter M controls the strength of regularisation: as $\sigma_{p_i} \to 0$, a_i shrinks smoothly to zero, and a flat low-variance neighbour collapses to its own mean rather than amplifying noise. We set M=5. The aligned past and future are $\tilde{\mathbf{p}}_i = a_i\mathbf{p}_i + b_i$ and $\tilde{\mathbf{f}}_i = a_i\mathbf{f}_i + b_i$, respectively.

3.2 Phase alignment

We shift the rescaled neighbour by the integer lag that maximises its sample cross-correlation with the query past:

$$\tau_i^* = \arg\max_{|\tau| \le S/4} \frac{1}{S - |\tau|} \sum_t \big(\mathbf{q}_{t+\tau} - \bar{\mathbf{q}}\big)\big(\tilde{\mathbf{p}}_{i,t} - \bar{\tilde{\mathbf{p}}}_i\big), \tag{2}$$

where $\bar{\mathbf{q}}$ and $\bar{\tilde{\mathbf{p}}}_i$ are the sample means. This is the maximum-likelihood lag estimator under stationary noise [Knapp and Carter, 1976]. Restricting $|\tau| \le S/4$ keeps the shift from consuming more than a quarter of the context. The same τ_i^* is applied to $\tilde{\mathbf{f}}_i$ to keep the past–future correspondence intact.

3.3 Auxiliary mechanics

Diversification. Top-K nearest neighbours in time-series corpora are typically near-duplicates (e.g., same day-of-week, same hour-of-day). We apply Maximal Marginal Relevance [Carbonell and Goldstein, 1998] with λ=0.3 to select K=10 neighbours from a retrieval pool of 20, maximising diversity among the chosen demonstrations while preserving relevance to the query.

Token-budget layout. Chronos-Bolt has a 2048-token budget; ten unmodified demonstrations of $S{+}H$=576 tokens overflow it. We pack a heterogeneous layout $M_d(S_d{+}H) + M_s(S_s{+}H) + S_q \le 2048$ with M_d=2, S_d=256, M_s=8, S_s=32. Two detailed demonstrations carry medium-range context; eight dense ones carry the most recent 32 steps and the full 64-step future. Futures are never truncated.

Future blend. The in-context prediction is the convex combination

$$\widehat{\mathbf{y}}_u \;=\; (1-\beta)\, f_\theta(C) + \beta \sum_{i=1}^{K} w_i\, \tilde{\mathbf{f}}_i, \qquad \beta{=}0.15, \tag{3}$$

where $w_i \propto \exp(-d_i/\tau)$ with $\tau = \mathrm{median}(\{d_i\})/5$ weights aligned futures by retrieval distance, C is the tiered context of the previous step, and β controls the contribution of the retrieval-based prior to the backbone forecast.

No-demonstration second pass. A second forward pass on the query alone, blended in the same way, gives $\widehat{\mathbf{y}}_c$. The final prediction averages the two, $\widehat{\mathbf{y}} = \alpha\widehat{\mathbf{y}}_u + (1-\alpha)\widehat{\mathbf{y}}_c$, with $\alpha{=}0.60$. Please note that we use this full configuration for the main headline numbers, but the central claim does not depend on these two post-processing terms: setting $\beta = 0$ and removing the consensus pass leaves a pure aligned-ICL variant whose only model-facing intervention is the aligned demonstration context. Table 7 shows that the pure aligned-ICL setting still outperforms TS-RAG in 6/7 datasets.

The full pseudocode is Algorithm 1. Hyperparameter values are listed in Appendix A; the same configuration is used throughout the paper.

Algorithm 1 Align-RAG.

Require: query past $\mathbf{q}$, retrieval pool $\mathcal{P} = \{(\mathbf{p}_i, \mathbf{f}_i, d_i)\}_{i=1}^{20}$, frozen TSFM f_θ

$\mathcal{S} \leftarrow \mathrm{MMR}(\mathbf{q}, \mathcal{P}, \lambda{=}0.3, K{=}10)$
for $(\mathbf{p}_i, \mathbf{f}_i, d_i) \in \mathcal{S}$ **do**
  $(a_i, b_i) \leftarrow$ Wiener affine fit of $\mathbf{p}_i$ to $\mathbf{q}$ (Eq. 1, $M{=}5$)
  $\tilde{\mathbf{p}}_i, \tilde{\mathbf{f}}_i \leftarrow a_i\mathbf{p}_i + b_i,\ a_i\mathbf{f}_i + b_i$
  Phase-align by integer-lag cross-correlation (Eq. 2)
end for
$C \leftarrow \textsc{TieredLayout}(\mathbf{q}, \{\tilde{\mathbf{p}}_i, \tilde{\mathbf{f}}_i\}_{i=1}^{K})$
$w_i \propto \exp(-d_i/\tau),\ \tau = \mathrm{median}(\{d_i\})/5$
$\widehat{\mathbf{y}}_u \leftarrow (1-\beta)\, f_\theta(C) + \beta \sum_i w_i\, \tilde{\mathbf{f}}_i$
$\widehat{\mathbf{y}}_c \leftarrow (1-\beta)\, f_\theta(\mathbf{q}) + \beta \sum_i w_i\, \tilde{\mathbf{f}}_i$
return $\widehat{\mathbf{y}} \leftarrow 0.60\,\widehat{\mathbf{y}}_u + 0.40\,\widehat{\mathbf{y}}_c$

4 Experiments

4.1 Setup

We evaluate on the seven datasets used by TS-RAG [Ning et al., 2025]: ETTh1, ETTh2, ETTm1, ETTm2, Weather, Exchange, and Electricity. Context length $S{=}512$ and horizon $H{=}64$ follow the TS-RAG protocol; longer-horizon results at $H \in \{96, 192\}$ are reported in Appendix G and preserve the gain on every (dataset, horizon) cell. Metrics are MSE and MAE. The frozen backbone for the main comparison is Chronos-Bolt-Base [Ansari et al., 2024]; cross-backbone results also use Chronos-2 [Ansari et al., 2025], TimesFM-2.0 [Das et al., 2024], Moirai [Woo et al., 2024], and Toto [Cohen et al., 2024]. For the head-to-head comparison against TS-RAG (Table 1) we use TS-RAG's own released top-20 retrieval index per (query, channel) so that any improvement isolates the effect of alignment from differences in retrieval quality. The TS-RAG numbers are produced by its released code and checkpoint and match the paper's reported values to $\leq 0.3\%$. Since Align-RAG uses labelled windows at inference time, we call it training-free adaptation rather than zero-shot forecasting; "zero-shot" denotes the backbone without demonstrations (Appendix J).

4.2 Main result on Chronos-Bolt

Table 1 reports MSE and MAE on the full test sets, together with paired moving-block bootstrap CIs on per-window squared-error differences (block 24, $N{=}2000$ resamples, Künsch, 1989, Liu and Singh, 1992).

Align-RAG attains the lower MSE on every benchmark and the lower MAE on six of seven. Five MSE wins exceed 2% relative reduction; five of six small-dataset bootstrap CIs lie strictly above zero.

Table 1: Align-RAG vs. TS-RAG [Ning et al., 2025] on a frozen Chronos-Bolt. Bold marks the lower of MSE/MAE per row. The right column is the paired moving-block bootstrap 95% CI on per-window MSE differences (TS-RAG − Align-RAG); positive means Align-RAG wins. Five of six small-dataset MSE wins are significant at α=0.05.

Dataset	n_{test}	Align-RAG	TS-RAG	Δ MSE / MAE	paired 95% CI
ETTh1	19,719	**0.3434 / 0.3612**	0.3556 / 0.3623	$-3.44\%/-0.30\%$	$[+0.0078, +0.0171]$
ETTh2	19,719	**0.2379 / 0.2964**	0.2451 / 0.2981	$-2.93\%/-0.57\%$	$[+0.0029, +0.0117]$
ETTm1	80,199	**0.2671 / 0.3047**	0.2904 / 0.3113	$-8.01\%/-2.12\%$	$[+0.0182, +0.0287]$
ETTm2	80,199	**0.1384 / 0.2188**	0.1465 / 0.2230	$-5.51\%/-1.90\%$	$[+0.0061, +0.0102]$
Weather	219,996	**0.1372 / 0.1729**	0.1453 / 0.1770	$-5.60\%/-2.34\%$	$[+0.0069, +0.0097]$
Exchange	11,632	**0.0620 / 0.1715**	0.0624 / 0.1716	$-0.69\%/-0.05\%$	$[-0.0015, +0.0025]$
Electricity	1,668,237	**0.1119** / 0.2018	0.1120 / 0.2003	$-0.07\%/+0.73\%$	—
Wins / avg				$7/7$ MSE (avg -3.75%); $6/7$ MAE (avg -0.93%)	

Chronos-Bolt inference is deterministic, so the only source of randomness in a paired comparison is the choice of test windows.

4.3 Alignment, not retrieval ranking, drives the improvement

We isolate the contribution of alignment from that of retrieval ranking with a 2×2 ablation on Chronos-Bolt (Table 2). The retriever is either TS-RAG's released top-20 index (*ranked*) or uniform random sampling from the leakage-safe train split (*random*). Alignment is either applied (*aligned*, the amplitude and phase steps of Section 3) or skipped (*no align*; retrieved windows enter the context unscaled and unshifted).

Table 2: Retriever × alignment ablation on Chronos-Bolt (MSE). "Aligned" applies steps (ii)+(iii) (amplitude and phase); "no align" skips both, leaving retrieved windows unscaled and unshifted. Δ is relative to zero-shot.

Variant	ETTh1	ETTh2	ETTm1	ETTm2	Weather	Exch.	Elec.	avg Δ
Zero-shot	0.3615	0.2516	0.3107	0.1486	0.1524	0.0671	0.1132	0%
Ranked retriever, aligned	0.3461	0.2462	**0.2900**	**0.1438**	0.1459	**0.0652**	**0.1108**	**−3.65%**
Ranked retriever, no align	0.3656	0.3131	0.3008	0.2121	0.1515	0.1271	0.1130	+22.12%
Random retriever, aligned	**0.3450**	**0.2437**	0.2929	0.1450	**0.1455**	0.0677	0.1130	−2.81%
Random retriever, no align	0.3535	0.2972	0.3015	0.2061	0.1508	0.1252	0.1178	+20.19%

With alignment fixed, retriever choice costs at most 0.84 pp average MSE: switching from the ranked retriever to uniform random windows still beats the trained TS-RAG mixer on $4/7$ datasets in absolute MSE, and on four datasets (ETTh1, ETTh2, Weather, Electricity) the random retriever is within 0.5 pp of the ranked one. Retriever quality matters more on Exchange and Electricity, the two non-periodic datasets where shape-matched retrieval carries channel-specific information. Even there, random retrieval is competitive.

With the retriever fixed, removing alignment under either choice regresses MSE by 20–22 pp versus zero-shot. We further investigate the mechanism through which alignment changes the model's behaviour in Section 5.

4.4 Cross-backbone generalisation

We apply the same method, with the same configuration, to five frozen TSFMs on the same seven datasets, at S=512, H=64. The retriever is rebuilt from scratch as a per-dataset, per-channel, train-split-only z-normalised Euclidean nearest-neighbour index, with a strict leakage audit.

All five backbones reduce MSE on average, ranging from -2.45% (Toto) to -13.73% (Moirai). Moirai, the backbone furthest from the Chronos lineage on which TS-RAG was trained, gains -13.73% on average and wins on every dataset, with the largest single cell a -29.95% MSE reduction on ETTm1 and -20.18% on Weather. TimesFM-2.0 wins on the four ETT datasets at up to -12.43% but regresses on Exchange ($+2.03\%$). Bolt and Chronos-2 win every benchmarked dataset

Table 3: Cross-backbone results: zero-shot MSE (ZS) and Align-RAG's relative MSE change vs. that backbone's zero-shot ($\Delta\%$); negative $\Delta\%$ means Align-RAG wins. Weather and Electricity were excluded from zero-shot evaluation for TimesFM in Ning et al. [2025]'s protocol on the grounds of pretraining-corpus overlap, so we omit those two cells for TimesFM-2.0 to remain consistent with their leakage treatment. All five backbones use the same in-house z-normalised Euclidean nearest-neighbour retriever; absolute MSE values therefore differ from Table 1's Align-RAG cells, which use TS-RAG's released top-20 index. The table evaluates untuned transfer, not uniform improvement in every cell (Appendix K).

Dataset	Bolt		Chronos-2		TimesFM		Moirai		Toto		row avg
	ZS	$\Delta\%$	ZS	$\Delta\%$	ZS	$\Delta\%$	ZS	$\Delta\%$	ZS	$\Delta\%$	$\Delta\%$
ETTh1	0.3046	$\mathbf{-4.89}$	0.3158	$\mathbf{-6.39}$	0.3165	$\mathbf{-7.10}$	0.3359	$\mathbf{-5.30}$	0.2931	$\mathbf{-4.01}$	-5.54
ETTh2	0.2278	$\mathbf{-0.27}$	0.2444	$\mathbf{-4.08}$	0.2451	$\mathbf{-1.12}$	0.2696	$\mathbf{-9.06}$	0.2217	$+0.04$	-2.89
ETTm1	0.2633	$\mathbf{-11.13}$	0.2645	$\mathbf{-11.65}$	0.2674	$\mathbf{-9.05}$	0.3962	$\mathbf{-29.95}$	0.2652	$\mathbf{-16.09}$	-15.58
ETTm2	0.1403	$\mathbf{-5.29}$	0.1408	$\mathbf{-0.13}$	0.1605	$\mathbf{-12.43}$	0.1822	$\mathbf{-16.60}$	0.1422	$+2.20$	-6.45
Weather	0.1521	$\mathbf{-6.26}$	0.1618	$\mathbf{-1.97}$	—	—	0.2132	$\mathbf{-20.18}$	0.1413	$\mathbf{-0.24}$	-7.16
Exch.	0.0651	$\mathbf{-1.20}$	0.0584	$\mathbf{-0.68}$	0.0636	$+2.03$	0.0599	$\mathbf{-1.24}$	0.0460	$+0.25$	-0.17
Elec.	0.1043	$+0.04$	0.1067	$\mathbf{-1.87}$	—	—	0.1628	$\mathbf{-13.82}$	0.1149	$+0.69$	-3.74
wins		6/7		7/7		4/5		7/7		3/7	
avg $\Delta\%$		-4.14		-3.83		-5.53		$\mathbf{-13.73}$		-2.45	-5.96

where they are evaluated, with Bolt's Electricity result ($+0.04\%$) effectively a tie. Toto is the most uneven case: it gains on ETTh1, ETTm1, and Weather but regresses by $+0.04\%$ to $+2.20\%$ on the other four datasets, and its -2.45% average is the smallest of the five backbones.

4.5 Component ablation

Table 4 reports a cumulative-add ablation on the five small datasets, with Exchange and Electricity reported separately under the table. Each row adds one component on top of the previous.

Table 4: Cumulative ablation on Chronos-Bolt, MSE change vs. TS-RAG. The last column is the average across the five smaller datasets; Electricity (n_{test}=1.67M) and Exchange (n_{test}=11.6k) span two orders of magnitude in test-set size and are reported separately under the table.

Variant	ETTh1	ETTh2	ETTm1	ETTm2	Weather	avg
Zero-shot Chronos-Bolt	$+1.7\%$	$+2.7\%$	$+7.0\%$	$+1.4\%$	$+4.9\%$	$+3.5\%$
+ CC align, K=2, β=0.1	-2.1%	$+1.6\%$	-4.1%	-1.1%	-5.5%	-2.2%
+ tiered layout, K=10, affine	-2.0%	-2.1%	-5.7%	-2.6%	-3.5%	-3.2%
+ MMR (λ=0.3) [steps i–v]	-2.9%	-2.1%	-7.4%	-4.0%	-5.4%	-4.4%
+ consensus pass (α=0.60) [full]	$\mathbf{-3.4\%}$	$\mathbf{-2.9\%}$	$\mathbf{-8.0\%}$	$\mathbf{-5.5\%}$	$\mathbf{-5.6\%}$	$\mathbf{-5.1\%}$

On Electricity, steps (i)–(v) regress to $+2.7\%$ MSE and the second pass (vi) recovers a tie at -0.07%; on Exchange the full method lands at -0.69%. The amplitude and phase steps alone close most of the gap to TS-RAG on the small datasets; the layout and diversification each contribute about a point of average MSE; the second-pass average contributes a further point and resolves the Electricity calibration regression that the in-context steps alone introduce on a 321-channel corpus (Appendix C). A leave-one-out variant on ETTh1 and ETTm1 confirms that removing any single component costs 0 to 2.8 pp MSE (Appendix B).

5 Aligned demonstrations elicit ridge-like behaviour

In this section, we examine the in-context behaviour of the frozen TSFM under our alignment, and ask whether the model's demonstration-induced prediction shift is consistent with the closed-form ridge prediction shift on the same $(\mathbf{p}_i, \mathbf{f}_i)$ pairs. The investigation combines a cosine signature relating the two prediction shifts and a future-shuffle causal control that distinguishes regression on the pairs from averaging over their futures.

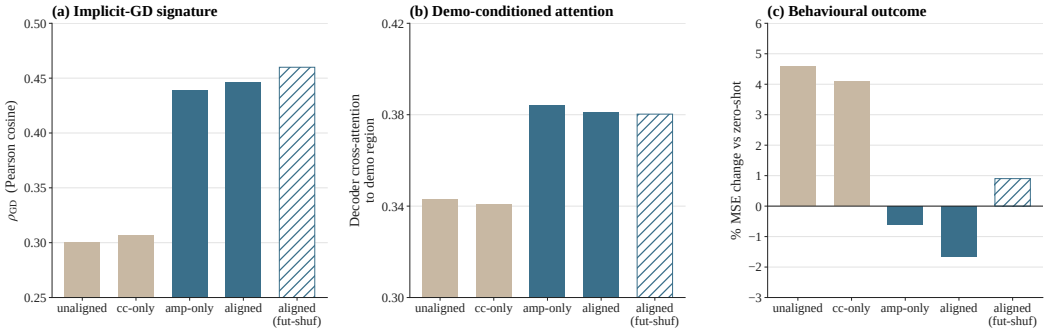


Figure 2: Mechanism evidence on Chronos-Bolt across five demonstration conditions (N=200 windows per cell, averaged over the seven datasets). (a) Implicit-GD signature ρ_{GD}: amplitude alignment lifts ρ_{GD} from 0.30 (`unaligned`, `cc_only`) to $\sim$0.45 (`amp_only`, `aligned`); the future-shuffle condition matches that level. (b) Decoder cross-attention to the demonstration region rises in the same conditions. (c) MSE change versus zero-shot: `amp_only` and `aligned` reduce MSE, whereas `aligned_fut_shuf` reverses the gain even though ρ_{GD} stays high in panel (a) – the model continues to track an implicit ridge solution, but on corrupted demonstration pairs.

Probe. Let $\widehat{\mathbf{y}}_{\mathrm{model}}$ and $\widehat{\mathbf{y}}_{\mathrm{zs}}$ denote the model's forecast with and without the aligned demonstrations, respectively. The closed-form ridge predictor is

$$\widehat{\mathbf{y}}_{\mathrm{ridge}} \;=\; \mathbf{q}^{\top}(\mathbf{P}^{\top}\mathbf{P} + \lambda \mathbf{I})^{-1}\mathbf{P}^{\top}\mathbf{F}, \qquad \lambda = 1,$$

where $\mathbf{P} \in \mathbb{R}^{K \times S}$ stacks the K=10 aligned demonstration pasts and $\mathbf{F} \in \mathbb{R}^{K \times H}$ stacks their aligned futures. The implicit-GD signature is the per-window cosine

$$\rho_{\mathrm{GD}}(w) \;=\; \cos\big(\widehat{\mathbf{y}}_{\mathrm{model}}(w) - \widehat{\mathbf{y}}_{\mathrm{zs}}(w),\; \widehat{\mathbf{y}}_{\mathrm{ridge}}(w) - \widehat{\mathbf{y}}_{\mathrm{zs}}(w)\big).$$

A pretrained TSFM also handles seasonality, trend, and recent-value extrapolation, so ρ_{GD}=1 is not reachable; the question is whether a regression-on-pairs component shows up above those other behaviours. We evaluate ρ_{GD} under five conditions per window: `unaligned` (raw retrievals), `amp_only`, `cc_only`, `aligned`, and `aligned_fut_shuf` (aligned demonstrations with futures permuted within the set, breaking the pair correspondence). Confidence intervals are computed by paired moving-block bootstrap with block length 20 and $N = 2000$ resamples over 200 paired test windows per dataset.

Result. ρ_{GD} rises from $[0.17, 0.48]$ unaligned (mean 0.30) to $[0.37, 0.59]$ aligned (mean 0.45). The paired bootstrap of $\rho_{\mathrm{GD}}^{\mathrm{aligned}} - \rho_{\mathrm{GD}}^{\mathrm{unaligned}}$ excludes zero on $7/7$ Chronos-Bolt datasets (smallest gap $+0.07$ on Electricity, largest $+0.25$ on Exchange). Cross-attention to the demonstration region rises on $6/7$ datasets (Fig. 2b). Of the two alignment components, amplitude carries the effect: `amp_only` matches `aligned` on $7/7$ datasets and `cc_only` reaches significance only on ETTh2.

Future-shuffle causal control. The shuffle permutes $\{\mathbf{f}_i\}$ within each demonstration set, leaving the pasts and the alignment intact. Under a futures-averaging account the shuffle is innocuous: the set of futures is unchanged. Under regression on the pairs it biases the closed-form ridge solution in a known direction, and a model that tracks that solution should inherit the bias. The latter is what we observe. The shuffle worsens MSE on $4/7$ datasets (mean $+2.6$ pp, median $+0.9$ pp); the largest reversal is on ETTm1, where the -4.2% gain from alignment flips to $+12.2\%$. The remaining three datasets, whose alignment-vs-unaligned gap is already small, move by less than 3 pp in either MSE or ρ_{GD}. Across all seven, ρ_{GD} remains high under the shuffle (mean 0.46, range $[0.34, 0.61]$): the model continues to implement an implicit ridge regression, now on the corrupted pairs $(\mathbf{p}_i, \mathbf{f}_i')$, and inherits whichever direction of bias that corruption induces. A futures-averaging account predicts neither the directional MSE shift nor the preservation of ρ_{GD}.

The probe replicates on Chronos-2's group-attention pathway with $\rho_{\mathrm{GD}}^{\mathrm{aligned}} > \rho_{\mathrm{GD}}^{\mathrm{unaligned}}$ on $6/7$ datasets ($13/14$ cells across the two backbones, Appendix D). Figure 3 shows one ETTm1 window: raw retrievals miss the query in both amplitude and phase; alignment collapses the demonstration cloud onto the query's regime, the model's prediction shift becomes visibly proportional to the closed-form ridge shift on the same pairs, and MSE drops from 2.06 (zero-shot) to 0.70 (aligned).

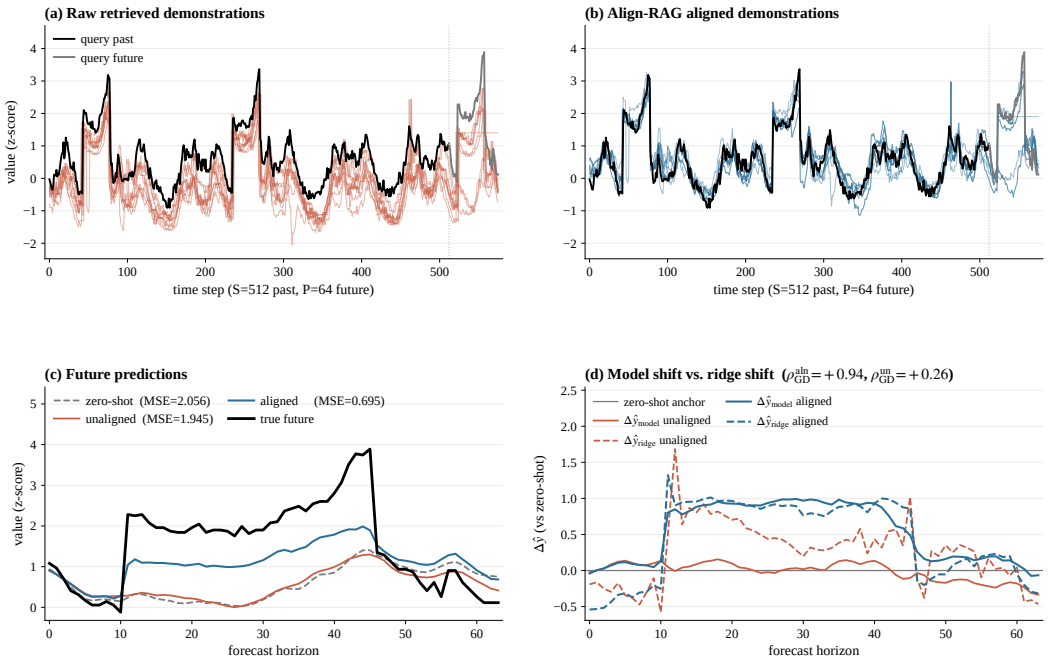


Figure 3: One ETTm1 query window. (a) Raw retrieved demonstrations have the wrong amplitude and phase relative to the query, and the resulting unaligned ICL prediction (MSE 1.94, $\rho_{\mathrm{GD}}{=}{+}0.26$) misses the true future. (b) Aligned demonstrations cluster around the query. (c) Aligned ICL recovers the trajectory (a $2.9\times$ reduction in MSE). (d) Under aligned, the model's prediction shift visibly tracks the closed-form ridge prediction shift on the same demos ($\rho_{\mathrm{GD}}{=}{+}0.94$); under unaligned, the two are nearly orthogonal ($\rho_{\mathrm{GD}}{=}{+}0.26$).

6 Conclusion

In this paper we revisited the assumption underlying recent retrieval-augmented forecasting on frozen Time Series Foundation Models: that learned fusion is required for a frozen backbone to benefit from retrieved context. We introduced Align-RAG, a training-free method that aligns retrieved past–future windows to the query in amplitude and phase before they enter the backbone's context. Under the standard seven-dataset TS-RAG benchmark protocol with a frozen Chronos-Bolt, Align-RAG matches or exceeds the trained TS-RAG mixer on every dataset in MSE, and the same configuration improves zero-shot performance on four additional frozen TSFMs without per-backbone tuning. A retriever $\times$ alignment ablation locates the gain in demonstration alignment rather than retrieval ranking, and a behavioural probe shows that under our alignment the model's prediction shift tracks the closed-form ridge prediction shift on the same demonstrations, with a future-shuffle causal control consistent with regression on the past–future pairs.

These findings carry empirical and mechanistic implications. Empirically, aligned demonstrations carry the bulk of the improvement in retrieval-augmented forecasting on a frozen TSFM, and learned fusion appears largely to compensate for the absence of such alignment. Mechanistically, frozen pretrained TSFMs admit a behavioural regression structure on aligned past–future pairs, consistent with predictions from the in-context learning literature on synthetic-regression-trained transformers. Natural future directions are richer retrieval geometries, including multivariate and covariate-informed settings.

Limitations. The mechanism evidence is correlational: $\rho_{\mathrm{GD}} = 0.45$ aligned versus 0.30 unaligned is directional evidence, not the equivalence ($\rho_{\mathrm{GD}}{=}1$) given by construction theorems. Our evaluation uses the standard univariate TS-RAG benchmark for direct comparison to the state-of-the-art; multivariate covariates and distribution shifts remain open.

Acknowledgments and Disclosure of Funding

Mohammad Asadi was supported by the Amazon AI PhD Fellowship and the Stanford HAI Graduate Fellowship during this work. We thank the Amazon fellowship program and Stanford HAI for their support. Soheil Hor, Bardiya Akhbari, Layne C. Price, and Raviteja Anantha are affiliated with Amazon; the authors declare no other competing interests.

References

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *International Conference on Learning Representations*, 2023.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. arXiv:2403.07815.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336, 1998.

Ben Cohen, Emaad Khwaja, Kan Wang, Charles Masson, Elise Ramé, Youssef Doubli, and Othmane Abou-Amal. Toto: Time series optimized transformer for observability. *arXiv preprint arXiv:2407.07874*, 2024.

Marco Cuturi and Mathieu Blondel. Soft-DTW: a differentiable loss function for time-series. In *International Conference on Machine Learning*, 2017.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning*, 2024. arXiv:2310.10688.

Matthew Faw, Rajat Sen, Yichen Zhou, and Abhimanyu Das. In-context fine-tuning for time-series foundation models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 16355–16374. PMLR, 2025. arXiv:2410.24087.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*, 2022.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan O. Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 21774–21797. PMLR, 2025.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020.

C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320–327, 1976. doi: 10.1109/TASSP.1976.1162830.

Hans R. Künsch. The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17(3):1217–1241, 1989.

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

Jingwei Liu, Ling Yang, Hongyan Li, and Shenda Hong. Retrieval-augmented diffusion models for time series forecasting. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2410.18712.

Regina Y. Liu and Kesar Singh. Moving blocks jackknife and bootstrap capture weak dependence. In Raoul LePage and Lynne Billard, editors, *Exploring the Limits of Bootstrap*, pages 225–248. Wiley, New York, 1992.

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. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2503.07649.

John Paparrizos and Luis Gravano. k-Shape: Efficient and accurate clustering of time series. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1855–1870, 2015.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 2023. arXiv:2302.00083.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: Retrieval-augmented black-box language models. In *NAACL*, 2024. arXiv:2301.12652.

Kutay Tire, Ege Onur Taga, Muhammed Emrullah Ildiz, and Samet Oymak. Retrieval augmented time series forecasting. 2024.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 35151–35174. PMLR, 2023.

Norbert Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*. MIT Press, 1949.

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

Silin Yang, Dong Wang, Haoqi Zheng, and Ruochun Jin. TimeRAG: Boosting LLM time series forecasting via retrieval-augmented generation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10889933. arXiv:2412.16643.

Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. Matrix profile I: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets. In *ICDM*, 2016.

Huanyu Zhang, Chang Xu, Yifan Zhang, Zhang Zhang, Liang Wang, and Jiang Bian. TimeRAF: Retrieval-augmented foundation model for zero-shot time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 37(9):5654–5665, 2025. doi: 10.1109/TKDE.2025.3579137.

Jiaping Zhao and Laurent Itti. shapeDTW: Shape dynamic time warping. *Pattern Recognition*, 2018.

A Implementation details

Retrieval index. Section 4.2 uses the released top-20 per (query, channel) index from TS-RAG [Ning et al., 2025]. Section 4.4 uses an in-house per-dataset, per-channel, train-split-only z-normalised Euclidean nearest-neighbor index, with a strict leakage audit that rejects any returned neighbor whose $S{+}H$-step span enters the validation or test split. Distances are Euclidean in the normalised-past space.

Hyperparameters. $K{=}10$, $\lambda_{\mathrm{MMR}}{=}0.3$, $\beta{=}0.15$, $\alpha{=}0.60$, $M{=}5$, $(M_d, S_d, M_s, S_s){=}(2, 256, 8, 32)$, CC search range $|\tau| \le S/4 = 128$, affine fit on $\mathbf{p}_i$ only. We use the same values for every dataset and backbone reported. Sensitivity sweeps over α (Table 6) and component leave-one-out (Table 5) are in the appendix.

Context layout. The tiered layout uses $2(256{+}64){+}8(32{+}64){+}512 = 1920$ of the 2048 positions available in Chronos-Bolt: two demonstrations keep 256 past steps, eight keep 32, all ten keep the full 64-step future, and the query's 512 steps are kept intact.

Compute. A single H100 evaluates Align-RAG with Chronos-Bolt across the seven datasets in under six wall-clock hours.

Code and reproducibility. The Align-RAG implementation is approximately 720 lines of Python across six files (alignment, retrieval interface, run-Bolt, run-Chronos-2, aggregation, metrics) and contains no learned parameters. Code, the cross-backbone retrieval-index build script, and the exact bootstrap-resample seeds are released alongside the paper. Section 4.2 reuses TS-RAG's public top-20 retrieval index without modification. The cross-backbone retrieval index of Section 4.4 uses a stride-1 window scan with z-normalisation per channel, Euclidean distance on the normalised past, and a leakage filter that removes any candidate whose $S{+}H$-step span overlaps the validation or test split.

B Leave-one-out component ablation

Table 5 removes each component in turn from the unified implementation on two representative datasets. Every component is non-trivial on at least one of the two datasets.

Table 5: Leave-one-out component ablation on Chronos-Bolt. MSE change vs. full Align-RAG on two datasets.

Removed	ETTh1		ETTm1	
	MSE	Δ	MSE	Δ
nothing (full Align-RAG)	0.3452	–	0.2688	–
− CC phase alignment	0.3451	$+0.0\%$	0.2701	$+0.5\%$
− MMR diversification	0.3484	$+0.9\%$	0.2738	$+1.9\%$
− tiered layout (use $K{=}2$)	0.3459	$+0.2\%$	0.2709	$+0.8\%$
− future blend ($\beta{=}0$)	0.3518	$+1.9\%$	0.2762	$+2.8\%$

C Second-pass averaging: decomposition and α sweep

MSE/MAE asymmetry. The two predictions $\widehat{\mathbf{y}}_u$ and $\widehat{\mathbf{y}}_c$ have per-window error correlation $r \in [0.94, 0.99]$ across the seven datasets, with 87–94% same-sign agreement. Averaging same-sign errors of comparable magnitude reduces tails more than it reduces medians: on Electricity, the per-step $q{=}0.99$ error quantile drops from $|e_u|{=}1.291$ and $|e_c|{=}1.270$ to 1.265 at the average, while the $q{=}0.50$ quantile moves by less than 10^{-3}. This is why the second pass helps MSE more than MAE.

Murphy–Theil decomposition of the Electricity penalty. Under steps (i)–(v) alone, Electricity is the single dataset on which Align-RAG loses $+3.1\%$ MSE relative to TS-RAG. Decomposing

the per-window MSE of $\widehat{\mathbf{y}}_u$ and $\widehat{\mathbf{y}}_c$ as $\mathrm{MSE} = \mathrm{bias}^2 + \mathrm{scale}^2 + \mathrm{shape}$ with $\mathrm{bias}^2 = (\bar{\widehat{y}} - \bar{y})^2$, $\mathrm{scale}^2 = (\sigma_{\widehat{y}} - \sigma_y)^2$, and $\mathrm{shape} = 2\sigma_{\widehat{y}}\sigma_y(1-\rho)$ gives:

	unified $\widehat{y}_u$	consensus $\widehat{y}_c$	gap (unified − cons.)
bias^2	0.03328	0.03181	+0.00147 (39% of gap)
scale^2	0.01833	0.01677	+0.00157 (42% of gap)
shape	0.06363	0.06290	+0.00072 (19% of gap)
MSE	0.11524	0.11148	+0.00376

81% of the in-context penalty lives in level and scale, not shape. The shape correlation ρ=0.64 is unchanged to 10^{-3} between the two predictions. Averaging at α=0.60 closes the gap to −0.07% MSE.

α sweep. Table 6 reports a fine sweep over $\widehat{\mathbf{y}}_{\text{final}} = \alpha\widehat{\mathbf{y}}_u + (1-\alpha)\widehat{\mathbf{y}}_c$. The MSE-win count is 7/7 for every $\alpha \in [0.50, 0.62]$, so the configuration is robust over a wide range of mixing weights. Within that range $\alpha \in \{0.60, 0.62\}$ also achieves 6/7 MAE wins; we pick α=0.60 once as the released setting and hold it fixed.

Table 6: Fine α sweep. Win counts out of seven and averages relative to TS-RAG.

α	MSE wins	MSE avg	MAE wins	MAE avg	elec. MSE	elec. MAE	exch. MSE	exch. MAE
0.50	7/7	−3.49%	5/7	−0.82%	−0.47%	+0.44%	−0.12%	+0.25%
0.58	7/7	−3.71%	5/7	−0.92%	−0.16%	+0.67%	−0.59%	+0.00%
0.60	**7/7**	**−3.75%**	**6/7**	**−0.93%**	**−0.07%**	**+0.73%**	**−0.69%**	**−0.05%**
0.62	6/7	−3.78%	6/7	−0.95%	+0.03%	+0.79%	−0.77%	−0.09%
0.70	6/7	−3.84%	6/7	−0.97%	+0.47%	+1.08%	−1.00%	−0.22%

D Mechanism: extended evidence

This appendix collects the per-dataset within-condition evidence behind the implicit-ridge mechanism summarised in Section 5: the within-dataset signature/MSE correlation, the cross-architecture replication on Chronos-2, and a concrete worked window.

Within-dataset GD-MSE correlation. Holding dataset fixed and correlating, across the five conditions, ρ_{GD} with the MSE improvement vs. zero-shot (Fig. 4): on 5/7 datasets the within-dataset Pearson correlation is large and positive (ETTh1 $r = +0.99$, ETTm2 +0.99, Weather +0.94, Electricity +0.98, ETTh2 +0.48). The two negative cases are diagnostic. ETTm1 ($r = -0.61$): the highest-ρ_{GD} cell is `aligned_fut_shuf` but its ridge target is biased by the shuffle, so the model faithfully implements that biased ridge and the MSE worsens. Exchange Rate ($r = -1.00$): zero-shot is already near the dataset's noise floor (MSE 0.062), so any implicit-ridge update moves the prediction off that floor. Both exceptions strengthen rather than weaken the mechanism: the signature measures *whether* the model is doing implicit ridge regression on the demos; the MSE outcome additionally requires the demos to carry useful information beyond zero-shot.

Cross-architecture replication. Figure 5 replicates the implicit-GD signature on Chronos-2, which has a fundamentally different ICL pathway (group attention over series rather than concatenation). On the same seven datasets (N=100 windows each), aligned demos again yield a higher signature than unaligned on 6/7 datasets (Weather is the single inversion). Combining both backbones, $\rho_{\mathrm{GD}}^{\mathrm{aligned}} > \rho_{\mathrm{GD}}^{\mathrm{unaligned}}$ on 13/14 (dataset, backbone) cells.

Concrete window. The ETTm1 window in Fig. 3 (main paper, test index 50998) illustrates the mechanism on a single query. The ten raw retrievals are misaligned in both amplitude and phase, and the resulting unaligned ICL prediction ($\rho_{\mathrm{GD}}^{\mathrm{un}}$=+0.26, MSE 1.94) is barely distinguishable from zero-shot (MSE 2.06). After alignment, the demonstration cloud collapses onto the query's regime, the model's prediction shift tracks the closed-form ridge shift on the same demos almost step-for-step ($\rho_{\mathrm{GD}}^{\mathrm{aln}}$=+0.94), and MSE falls to 0.70 – a 2.9× reduction over zero-shot on this window.

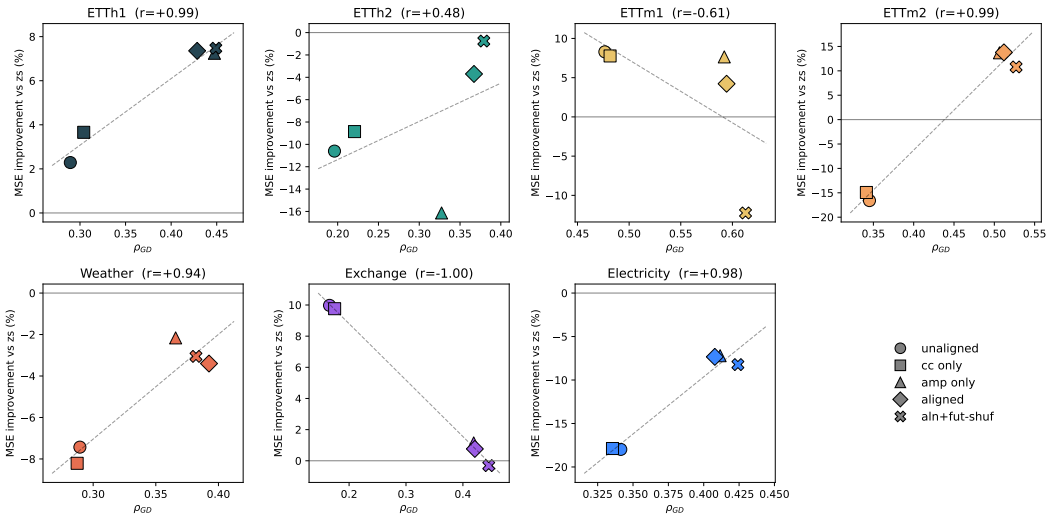


Figure 4: Within-dataset correlation between implicit-GD signature ρ_{GD} and MSE improvement vs. zero-shot (%). One panel per dataset, five markers per panel (one per condition); Pearson r shown. Five datasets (ETTh1, ETTh2, ETTm2, Weather, Electricity) show large positive r, as the mechanism predicts. ETTm1 ($r = -0.61$) is a diagnostic case (the `aligned_fut_shuf` cell has high ρ_{GD} but a biased ridge target, so the model's faithful tracking yields worse MSE; see future-shuffle control). Exchange ($r = -1.00$) is at zero-shot's noise floor (MSE 0.062), where any in-context shift moves prediction off the floor; we discuss both anomalies in Sec. 5.

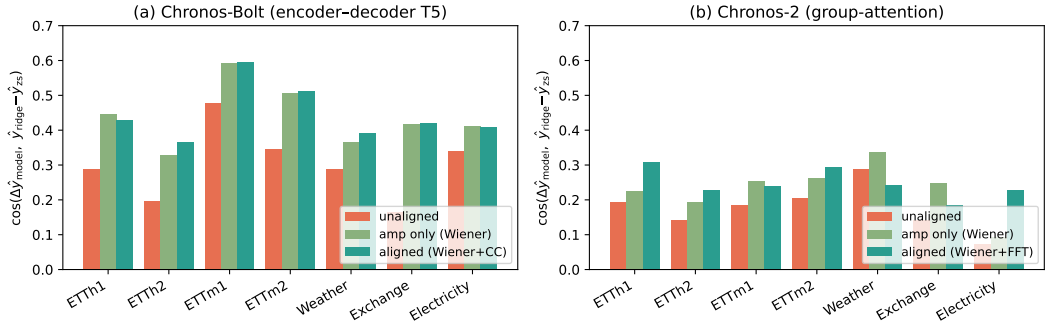


Figure 5: Cross-architecture replication of the implicit-GD signature. (a) Chronos-Bolt (T5 encoder–decoder, CC phase alignment; N=200): aligned > unaligned on $7/7$, all bootstrap-significant. (b) Chronos-2 (group attention, FFT phase alignment; N=100): aligned > unaligned on $6/7$. The `amp_only` condition is competitive with `aligned` on Chronos-2, matching the cross-backbone Wiener-vs-clip ablation in Sec. 4.4: FFT phase transfer is most useful on the ETT family.

E Isolating the in-context contribution from blend and second pass

Align-RAG combines four in-context steps (diversification, amplitude alignment, phase alignment, layout) with two post-processing additions: the future blend at β=0.15 and the second-pass average at α=0.60. Setting β=0 and/or α=0 isolates the contribution of the in-context steps alone. Table 7 reports the four corners of this 2×2 on Chronos-Bolt against TS-RAG.

Aligned demos fed through nothing but the in-context layout beat trained TS-RAG on $6/7$ datasets at average -1.64% MSE; the only exception is Electricity. The β blend recovers Electricity ($+4.1\% \rightarrow +1.0\%$); the α second pass closes the gap to -0.1%.

Table 7: Isolating the four in-context steps from the post-processing β blend and α consensus, on Chronos-Bolt. Each cell is per-dataset MSE change vs. TS-RAG; bold indicates a gain. Wins/7 counts datasets with strictly negative Δ MSE; avg $\Delta\%$ averages the seven per-dataset deltas.

Variant	ETTh1	ETTh2	ETTm1	ETTm2	Weather	Exchange	Electricity	wins	avg $\Delta\%$
Full Align-RAG (β=0.15, α=0.60)	-3.4%	-2.9%	-8.0%	-5.5%	-5.6%	-0.7%	-0.1%	$\mathbf{7/7}$	$\mathbf{-3.75\%}$
$-\alpha$ only (β=0.15, α=0)	-3.3%	-2.2%	-8.0%	-4.5%	-6.1%	-1.0%	$+2.7\%$	$6/7$	-3.20%
$-\beta$ only (β=0, α=0.60)	-1.0%	-1.5%	-4.2%	-4.8%	-2.6%	-2.5%	$+1.0\%$	$6/7$	-2.20%
$-\alpha$ and $-\beta$ (pure aligned ICL)	-1.0%	-0.9%	-4.9%	-3.2%	-3.3%	-2.2%	$+4.1\%$	$6/7$	-1.64%

F Wiener vs. clipped affine on cross-backbone Weather and Exchange

The plug-in scale ratio σ_q/σ_{p_i} produces Q_{99} values of 1.51 on ETTh1 but 3.0 on Weather and 6.7 on Exchange, with maxima exceeding 10^6 on the rare flat retrievals that any nearest-neighbor index occasionally returns. Two regularisers stabilise this tail: hard-clipping $a_i \in [1/M, M]$ at M=5, or Wiener shrinkage. On Weather, where the tail is heaviest, Wiener strictly dominates clipping on every backbone (Table 8); on Exchange the two regularisers are competitive, with Wiener slightly better on three of five backbones and clipping slightly better on the other two. On the five datasets where the tail is bounded, the two regularisers agree to four decimal places. We deploy Wiener shrinkage globally because it never explodes and matches or beats clipping on the two tail-heavy datasets.

Table 8: Step-(ii) regularisers on the two datasets where the plug-in scale ratio is unbounded. MSE change vs. each backbone's zero-shot. Underline marks the better regulariser. The Wiener row reproduces the deployed values from Table 3; the plug-in and clipped rows are from the same retrieval setup with only step (ii) substituted.

	Bolt	Chronos-2	TimesFM	Moirai	Toto
			Weather		
Plug-in	$+8.0\%$	$+10.2\%$	—	$+2.8\%$	$+11.7\%$
Clipped	$+1.3\%$	$+3.0\%$	—	$+0.3\%$	$+4.1\%$
Wiener	$\underline{-6.26\%}$	$\underline{-1.97\%}$	—	$\underline{-20.18\%}$	$\underline{-0.24\%}$
			Exchange		
Plug-in	$+3.0\%$	$+6.1\%$	$+4.8\%$	$+7.1\%$	$+8.0\%$
Clipped	$\underline{-2.9\%}$	-0.6%	$\underline{-1.4\%}$	$+3.6\%$	$+0.7\%$
Wiener	-1.20%	$\underline{-0.68\%}$	$+2.03\%$	$\underline{-1.24\%}$	$\underline{+0.25\%}$

G Multi-horizon results

The TS-RAG protocol of Ning et al. [2025] fixes the forecast horizon at H=64. To check that Align-RAG's gains are not specific to that one horizon, we evaluate at $H \in \{64, 96, 192\}$ on the five small datasets, with the same frozen Chronos-Bolt backbone and the same alignment hyperparameters as the main result (Section 4.2). The retrieval index here is the in-house z-normalised Euclidean nearest-neighbour index used in the cross-backbone evaluation (Section 4.4) rather than the TS-RAG released top-20 index used in Table 1; the two retrievers produce different absolute MSE values on the same backbone, so the H=64 column of Table 9 is not numerically identical to the corresponding cells of Table 1, even though the pipeline is otherwise the same.

The MSE gains hold at every (dataset, horizon) cell, ranging from -1.20% to -6.59%. ETTm1 has the largest gains ($\approx -6.5\%$ across horizons); ETTh2 and Exchange have the smallest. The MAE picture is similar but with smaller magnitudes (MAE is dominated by trend and shape rather than scale, which is where in-context demonstrations help most), with two cells where MAE is essentially flat ($+0.04\%$ on ETTm2 at H=64 and $+0.47\%$ on ETTh2 at H=64). The horizon ranking of the relative gains is monotone or nearly so on every dataset, so the H=64 result reported in the main paper is not a peculiarity of that specific horizon.

Table 9: Multi-horizon MSE on a frozen Chronos-Bolt. $\Delta\%$ is the relative MSE change vs. zero-shot at the same horizon; negative means Align-RAG wins.

Dataset	H	Zero-shot	Align-RAG	$\Delta\%$
ETTh1	64	0.3615	**0.3461**	-4.26%
ETTh1	96	0.3828	**0.3663**	-4.31%
ETTh1	192	0.4364	**0.4148**	-4.94%
ETTh2	64	0.2516	**0.2462**	-2.14%
ETTh2	96	0.2875	**0.2807**	-2.35%
ETTh2	192	0.3539	**0.3403**	-3.85%
ETTm1	64	0.3107	**0.2904**	-6.53%
ETTm1	96	0.3308	**0.3090**	-6.59%
ETTm1	192	0.3836	**0.3594**	-6.31%
ETTm2	64	0.1486	**0.1441**	-2.98%
ETTm2	96	0.1767	**0.1710**	-3.25%
ETTm2	192	0.2476	**0.2389**	-3.54%
Exchange	64	0.0671	**0.0652**	-2.90%
Exchange	96	0.0992	**0.0970**	-2.22%
Exchange	192	0.1938	**0.1914**	-1.20%

Table 10: Multi-horizon MAE on a frozen Chronos-Bolt. Each cell reports zero-shot / Align-RAG / $\Delta\%$ at the indicated horizon.

Dataset	H=64	H=96	H=192
ETTh1	$0.3650/\mathbf{0.3635}/-0.40\%$	$0.3795/\mathbf{0.3772}/-0.62\%$	$0.4117/\mathbf{0.4070}/-1.14\%$
ETTh2	$0.2992/0.3006/+0.47\%$	$0.3250/0.3257/+0.22\%$	$0.3688/\mathbf{0.3672}/-0.43\%$
ETTm1	$0.3183/\mathbf{0.3137}/-1.44\%$	$0.3331/\mathbf{0.3275}/-1.70\%$	$0.3653/\mathbf{0.3584}/-1.89\%$
ETTm2	$0.2235/0.2236/+0.04\%$	$0.2447/\mathbf{0.2439}/-0.34\%$	$0.2932/\mathbf{0.2911}/-0.72\%$
Exchange	$0.1778/\mathbf{0.1762}/-0.93\%$	$0.2190/\mathbf{0.2163}/-1.24\%$	$0.3138/\mathbf{0.3098}/-1.28\%$

H Comparison against RAFT

RAFT [Han et al., 2025] is a trained-adapter retrieval-augmented forecaster: a projection module and prediction head are trained on top of the retrieved patterns, on a frozen backbone. We compare Align-RAG against RAFT on the four ETT datasets at the two horizons (H=96, 192) reported by RAFT, with both methods using a frozen Chronos-Bolt backbone.

Table 11: Align-RAG vs. RAFT [Han et al., 2025] on a frozen Chronos-Bolt at $H \in \{96, 192\}$. Align-RAG attains the lower MSE on every cell. "ARvs. RAFT" is the relative MSE change of Align-RAG against RAFT; "ARvs. ZS" is the relative change against zero-shot Chronos-Bolt at the same horizon.

Dataset	H	RAFT	Bolt-ZS	Align-RAG	ARvs. RAFT	ARvs. ZS
ETTh1	96	0.3868	0.3828	**0.3663**	-5.29%	-4.31%
ETTh1	192	0.4225	0.4364	**0.4148**	-1.82%	-4.94%
ETTh2	96	0.2960	0.2875	**0.2807**	-5.17%	-2.35%
ETTh2	192	0.3840	0.3539	**0.3403**	-11.37%	-3.85%
ETTm1	96	0.3290	0.3308	**0.3090**	-6.08%	-6.59%
ETTm1	192	0.3634	0.3836	**0.3594**	-1.11%	-6.31%
ETTm2	96	0.1773	0.1767	**0.1710**	-3.59%	-3.25%
ETTm2	192	0.2425	0.2476	**0.2389**	-1.50%	-3.54%
Wins / avg					$8/8, -4.5\%$	$8/8, -4.4\%$

Align-RAG attains the lower MSE on $8/8$ cells with an average reduction of -4.5% relative to RAFT. The largest single gain is -11.37% on ETTh2 at H=192; the smallest is -1.11% on ETTm1 at H=192. Align-RAG also continues to improve on zero-shot Chronos-Bolt at every cell (-2.35% to -6.59%), so the comparison cannot be reduced to "RAFT is a weak baseline" – RAFT is a non-trivial

gain over zero-shot on six of eight cells, and Align-RAG still beats it on every one without training any parameters.

I Cross-dataset retrieval

A natural concern with retrieval-augmented forecasting is whether the retrieval pool needs to come from the same dataset as the query. We test the cross-dataset (xds) regime by replacing each target's same-corpus retrieval index with one drawn from a held-out ETT dataset (a different ETT subset than the target), keeping every other component of Align-RAG fixed. The two non-ETT targets, Exchange and Weather, retrieve only from the other datasets.

Table 12: Cross-dataset retrieval on a frozen Chronos-Bolt at $H{=}64$. "Same-corpus" draws retrievals from the target's own training split; "cross-dataset" draws them from a held-out ETT subset (for Exchange and Weather, from the other datasets). "xds vs. ZS" is the relative MSE change of cross-dataset Align-RAG vs. zero-shot; "xds vs. same-corpus" is the relative change vs. same-corpus Align-RAG (positive means same-corpus wins).

Target	Bolt-ZS	AR(same-corpus)	AR(cross-dataset)	xds vs. ZS	xds vs. same-corpus
ETTh1	0.3615	0.3434	**0.3418**	-5.44%	-0.46%
ETTh2	0.2516	**0.2379**	0.2487	-1.17%	$+4.52\%$
ETTm1	0.3107	**0.2671**	0.3036	-2.27%	$+13.65\%$
ETTm2	0.1486	**0.1384**	0.1427	-3.98%	$+3.06\%$
Exchange	0.0671	**0.0620**	0.0670	-0.2%	—
Weather	0.1524	**0.1372**	0.1540	$+1.1\%$	—

Cross-dataset Align-RAG still beats zero-shot on all four ETT targets (-1.17% to -5.44% MSE), so the gain does not vanish when retrieval shifts to an out-of-distribution corpus; of the two non-ETT targets, Exchange improves slightly (-0.2%) and Weather does not ($+1.1\%$). On ETTh1 the cross-dataset variant is essentially indistinguishable from the same-corpus variant (-0.46%); on ETTh2 and ETTm2 the same-corpus variant is meaningfully better ($+3$ to $+5$ percentage points); on ETTm1 the gap is largest ($+13.65\%$). The takeaway is two-sided: alignment of retrieved demonstrations is sufficient to extract a non-trivial gain even from cross-dataset retrieval, but the same-corpus regime is the operating point we recommend when in-domain retrieval is feasible.

J Evaluation settings

We distinguish four settings: (i) standard zero-shot inference, where the frozen backbone sees only the query past; (ii) retrieval within the target dataset, the setting of our main experiments; (iii) cross-dataset retrieval, where the pool excludes the target dataset (Appendix I); and (iv) parameter-free adaptation, where no parameters are trained, fine-tuned, or added. Align-RAG is a parameter-free adaptation used with (ii) or (iii).

K When are the gains small?

Alignment removes amplitude and phase mismatch but cannot make an uninformative retrieval predictive, so the gain depends on how useful retrieval is for a dataset and on the backbone's remaining headroom. Across datasets, our gains are rank-correlated with TS-RAG's published gains ($\rho{=}0.79$, $p{=}0.036$); across backbones, stronger zero-shot performance (e.g., Toto) goes with smaller gains, a pattern TS-RAG also reports [Ning et al., 2025].

L Retriever $\times$ alignment ablation on additional backbones

Table 13 repeats the ablation of Table 2 on three further backbones, and Table 14 decomposes each $2{\times}2$ into its main effects. On every backbone, the gain comes mainly from alignment.

Table 13: Retriever × alignment ablation (MSE) against each backbone's own zero-shot. Electricity is omitted; Weather is omitted for TimesFM-2.0 as in Table 3.

Backbone	Variant	ETTh1	ETTh2	ETTm1	ETTm2	Weather	Exch.	avg Δ
Chronos-2	zero-shot	0.3771	0.2659	0.3092	0.1487	0.1427	0.0596	0%
	ranked, aligned	0.3592	0.2542	0.2882	0.1430	0.1385	0.0592	−3.90%
	ranked, no align	0.3792	0.3224	0.2997	0.2112	0.1452	0.1228	+28.10%
	random, aligned	0.3571	0.2520	0.2888	0.1436	0.1376	0.0609	−3.65%
	random, no align	0.3666	0.3063	0.2974	0.2030	0.1451	0.1213	+25.05%
TimesFM-2.0	zero-shot	0.3790	0.2639	0.3116	0.1682	—	0.0646	0%
	ranked, aligned	0.3632	0.2568	0.2929	0.1602	—	0.0637	−3.81%
	ranked, no align	0.3836	0.3200	0.3044	0.2282	—	0.1323	+32.10%
	random, aligned	0.3636	0.2563	0.2971	0.1635	—	0.0663	−2.37%
	random, no align	0.3749	0.3036	0.3066	0.2244	—	0.1321	+30.06%
Moirai	zero-shot	0.3862	0.2824	0.4717	0.1900	0.1802	0.0583	0%
	ranked, aligned	0.3832	0.2772	0.4269	0.1815	0.1724	0.0604	−2.89%
	ranked, no align	0.3995	0.3354	0.4389	0.2446	0.1778	0.1267	+26.64%
	random, aligned	0.3878	0.2765	0.4455	0.1828	0.1710	0.0627	−1.44%
	random, no align	0.3976	0.3192	0.4618	0.2383	0.1772	0.1262	+25.67%

Table 14: Main effects of the ablation (percentage points of average MSE change vs. zero-shot). Chronos-Bolt is from Table 2.

	Chronos-Bolt	Chronos-2	TimesFM-2.0	Moirai
Alignment effect (ranked retriever)	+25.77	+32.00	+35.91	+29.54
Alignment effect (random retriever)	+23.00	+28.71	+32.43	+27.11
Retriever effect (aligned)	+0.84	+0.24	+1.43	+1.45

M Scope of the affine amplitude map

The amplitude step assumes that the scale and offset estimated from the retrieved and query pasts remain approximately appropriate over the forecast horizon. This is reasonable when the retrieved pair is a recurring pattern from the same local regime, but may fail under a future-side level or variance change, or a structural break, that is not present in the retrieved pair; such unmatched regime changes are a limitation of this setting, although not one specific to Align-RAG. The map only places the retrieved pair at the query's level and scale, and the forecasting itself is done by the frozen TSFM. An oracle that fits the map on the full past+future span changes MSE by only $+0.17\%$ on average and by at most 1.47% on any dataset.

N Sensitivity to the cross-correlation lag bound

The bound $|\tau| \leq S/4$ is a fixed design choice. Sweeping it over an $8\times$ range on Chronos-Bolt (Table 15) changes MSE by at most 0.12 percentage points: the retriever already returns nearly in-phase candidates (median selected $|\tau|$ of 0 or 1 sample), so cross-correlation is only a local correction. With a phase-invariant retriever, or when relevant offsets often exceed $S/4$, the bound may need to scale with the expected seasonality.

Table 15: Maximum-lag sweep on Chronos-Bolt: MSE change vs. zero-shot.

Maximum lag	32	64	128 (released)	256
ETTh1	−4.38%	−4.30%	−4.26%	−4.26%
ETTh2	−2.13%	−2.13%	−2.14%	−2.14%
ETTm1	−6.67%	−6.67%	−6.66%	−6.66%
ETTm2	−3.23%	−3.23%	−3.23%	−3.23%
Mean	−4.10%	−4.08%	−4.07%	−4.07%

O Oracle-future calibration of ρ_{GD}

On the windows of Section 5, we keep the aligned demonstration pasts, replace all demonstration futures with the query's true future, and recompute ρ_{GD} (Table 16). Aligned demonstrations close 56% of the unaligned-to-oracle gap. As ρ_{GD} is a directional cosine, the oracle is a calibration rather than a per-dataset upper bound (Exchange is non-monotone).

Table 16: Oracle-future calibration of ρ_{GD} on Chronos-Bolt. Gap closed $=$ (aligned $-$ unaligned)/(oracle $-$ unaligned), with 95% CI.

Dataset	Unaligned	Aligned	Oracle future	Gap closed [95% CI]
ETTh1	0.289	0.429	0.582	48% $[32, 62]$
ETTh2	0.196	0.367	0.579	45% $[25, 61]$
ETTm1	0.477	0.594	0.734	46% $[27, 68]$
ETTm2	0.345	0.512	0.665	52% $[35, 66]$
Weather	0.289	0.392	0.517	45% $[15, 85]$
Exchange	0.166	0.421	0.369	126% $[82, 181]$
Electricity	0.341	0.408	0.479	48% $[27, 75]$
Mean	0.301	0.446	0.561	56%

P Electricity

The -0.07% on Electricity in Table 1 is relative to TS-RAG; relative to zero-shot, Align-RAG improves MSE by 1.15% and TS-RAG by 1.06%. Electricity's retrieved windows are already close to the query before alignment (mean relative L_2 distance 0.61 vs. 1.13–2.78 on the other datasets; mean correlation 0.88 vs. 0.55–0.79), leaving little for an affine map to fix.

Q Composition with trained modules and a multivariate probe

Align-RAG does not touch the backbone's weights, so it composes with trained modules. Adding aligned demonstrations to the encoder context of TS-RAG's released model, with its retrieval and mixer unchanged, lowers MSE on all four ETT datasets (Table 17); unaligned, the same demonstrations leave ETTh1 MSE unchanged at 0.3555. On Chronos-2's learned group-attention (multivariate) pathway, with neighbouring windows as proxy variables, alignment improves over both zero-shot and the native in-context mechanism (Table 18). Both modules instance-normalise each retrieved window in isolation, which cancels a positive affine map, so aligned demonstrations must enter through the encoder context.

Table 17: Aligned demonstrations on top of TS-RAG's trained mixer (ARM), full test sets (MSE).

Dataset	TS-RAG (ARM only)	ARM + aligned demonstrations	Δ
ETTh1	0.3555	0.3507	-1.35%
ETTh2	0.2451	0.2419	-1.31%
ETTm1	0.2903	0.2728	-6.03%
ETTm2	0.1465	0.1454	-0.75%

Table 18: Align-RAG on Chronos-2's native in-context (multivariate) pathway; zero-shot is univariate; 500-window probes (MSE).

Dataset	Zero-shot	Native ICL	+ Align-RAG	vs. zero-shot	vs. native ICL
ETTh1	0.3565	0.3478	0.3332	-6.5%	-4.2%
ETTh2	0.2792	0.2633	0.2562	-8.2%	-2.7%

R Computational cost

Every step is constant or linear in the number of retrieved candidates K and the batch size B (Table 19; n: encoder tokens, $\approx$context$/16$; d=768; S: context length; P: horizon; L_{demo}: demonstration length). Table 20 profiles Chronos-Bolt; "optimised" adds FFT-based alignment and overlaps the two forward passes.

Table 19: Asymptotic cost of each step.

Step	Cost	in K	in S	in B
Retrieval	$B\,K\,(S{+}P)$ gathered	linear	linear	linear
Alignment	$B\,K\,S\log S$ (FFT over lags)	linear	$S\log S$	linear
Longer input context	$n\,d^2$, $n=(S+KL_{\text{demo}})/16$	linear	linear	linear
Two TSFM forwards	$n_1 d^2 + n_2 d^2$, n_2=33 fixed	constant	—	linear
Future blending	$B\,K\,P$ weighted mean	linear	—	linear

Table 20: Per-window latency (ms), peak GPU memory, and trained parameters on Chronos-Bolt.

	Zero-shot	TS-RAG	Align-RAG	Align-RAG (optimised)
Retrieval	—	0.021	0.021	0.021
Candidate selection (MMR 20→10)	—	—	0.017	0.017
Alignment (amplitude + phase)	—	—	0.088	0.002
TSFM forward	0.087	0.097	0.250	0.185
Blend with retrieved futures	—	0.000	0.004	0.004
Data movement	0.013	0.018	0.008	0.008
Total per window	0.100	0.136	0.388	0.237
Forward passes	1	1 (+ARM)	2	2
Encoder tokens	33	33	121+33	121+33
Peak GPU memory (MiB)	789	1,668	1,722	$\approx$ 2,040
Trained parameters	0	4,775,425	0	0
MSE	0.3615	0.3556	0.3426	0.3426
vs. TS-RAG, per window	—	—	2.9×	1.7×

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract claims (i) training-free closed-form alignment, (ii) outperforming the trained TS-RAG mixer on $7/7$ datasets in MSE on a frozen Chronos-Bolt with average -3.75% MSE, (iii) cross-backbone transfer to four further frozen TSFMs, and (iv) a behavioural mechanism analysis. Each is supported by Sections 3, 4.2, 4.4, and 5, respectively.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated "Limitations" paragraph at the end of Section 6 acknowledges that (i) the mechanism evidence is correlational rather than the equivalence the construction theorems give, (ii) the layout and second-pass steps run only on the Chronos family because patch-encoder backbones do not admit token-level context manipulation, and (iii) the retriever-$\times$-alignment ablation is reported on Chronos-Bolt in the main text and on three further backbones in Appendix L.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not introduce new theorems. Section 5 compares the model's prediction shift to the closed-form ridge prediction shift on the same demonstrations as a behavioural probe, with the closed-form expression for $\widehat{\mathbf{y}}_{\mathrm{ridge}}$ given in full and its connection to the in-context-learning literature cited.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Algorithm 1 gives the full pseudocode. All hyperparameters (K, λ_{MMR}, β, α, M, the tiered-layout shape (M_d, S_d, M_s, S_s), the cross-correlation search range, and the moving-block bootstrap configuration) are specified in Section 3 and Appendix A. We use the same values across every dataset and backbone reported. The benchmark protocol (datasets, $S{=}512$, $H{=}64$, metrics, retrieval index source) follows Ning et al. [2025].

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All seven evaluation datasets (ETTh1/h2, ETTm1/m2, Weather, Exchange, Electricity) are public benchmarks with the splits used by Ning et al. [2025]. All five backbones are publicly released frozen checkpoints. Align-RAG itself is approximately 720 lines of Python and contains no learned parameters; the implementation, the cross-backbone retrieval-index build script, and the exact bootstrap-resample seeds will be released alongside the camera-ready paper. The Section 4.2 comparison reuses TS-RAG's public top-20 retrieval index without modification. The code is available at `https://github.com/masadi-99/align-rag`

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Align-RAG is training-free, so there is no optimizer or training schedule to specify. Test-set splits, context length $S{=}512$, horizon $H{=}64$, retrieval-index construction (per-dataset, per-channel, z-normalised Euclidean nearest-neighbour with leakage filter), and all hyperparameters are reported in Sections 3, 4, and Appendix A. Hyperparameter sensitivity is documented in Appendix C (α sweep) and Appendix B (component leave-one-out).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 1 reports paired moving-block bootstrap 95% CIs on per-window squared-error differences (block length 24, N=2,000 resamples; Künsch, 1989, Liu and Singh, 1992). The mechanism analysis in Section 5 reports the same paired bootstrap on $\rho_{\mathrm{GD}}^{\mathrm{aligned}} - \rho_{\mathrm{GD}}^{\mathrm{unaligned}}$ over 200 paired test windows per dataset. The randomness source is window choice; Chronos-Bolt inference is deterministic.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix A reports that a single H100 GPU evaluates Align-RAG with Chronos-Bolt across all seven datasets in under six wall-clock hours. Cross-backbone runs use the same hardware. There is no training compute (the method is training-free).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses only publicly released frozen forecasting models and standard public time-series benchmarks (no human subjects, no scraped personal data, no novel data collection). The contribution is an algorithmic procedure layered on top of existing models.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper presents a training-free, model-agnostic algorithmic procedure for retrieval-augmented forecasting on top of frozen pretrained models. It introduces no new dataset, no new model weights, and no new release with novel deployment risks beyond those already attached to the pretrained backbones it consumes. Improvements in forecasting accuracy carry the same generic dual-use considerations as forecasting research broadly and we do not see a direct path to specific harms.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

Justification: The paper releases neither a new dataset nor a new model. The released artifact is a small training-free Python codebase that wraps existing public time-series foundation models; no high-risk asset is created.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Every backbone (Chronos-Bolt, Chronos-2, TimesFM-2.0, Moirai, Toto), every dataset (ETT family, Weather, Exchange, Electricity), every comparison method (TS-RAG, TimeRAF, RATD, RAFT, TimeRAG, RAF, TimesFM-ICF), and every supporting algorithm (MMR, moving-block bootstrap, soft-DTW, k-Shape, matrix profile) is cited to its original source. We use the public Chronos, Moirai, TimesFM, and Toto checkpoints under their published licenses, and the TS-RAG public top-20 retrieval index under its released terms.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The Align-RAG implementation (approximately 720 lines of Python across six files: alignment, retrieval interface, run-Bolt, run-Chronos-2, aggregation, metrics) and the cross-backbone retrieval-index build script will be released with documentation, the exact configuration used in every reported table, and the bootstrap-resample seeds. No model weights are introduced (the method is training-free).

Guidelines:

- The answer [N/A] means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper involves no crowdsourcing and no human subjects. All evaluation is on public time-series benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper involves no human subjects, so no IRB review applies.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Large language models are not part of Align-RAG's core methodology. The method operates on time-series foundation models (frozen Chronos-Bolt, Chronos-2, TimesFM-2.0, Moirai, Toto), none of which are language models, and applies closed-form numerical transforms (amplitude rescaling and integer-lag cross-correlation) to retrieved past–future windows. Any LLM use was confined to writing/editing assistance, which the policy exempts from declaration.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.