Retrieval Mechanisms Surpass Long-Context Scaling in Time Series Forecasting

Rishi Ahuja[1] Kumar Prateek[1] Simranjit Singh[1] Vijay Kumar[1]

[1]Department of Information Technology, Dr. B.R. Ambedkar National Institute of Technology Jalandhar

International Conference on Learning Representations (ICLR) 2026 - Time Series in the Age of Large Models (TSALM) Workshop

Motivation

Time Series Foundation Models borrow the **"Long Context"** paradigm from NLP: more history ⇒ better forecasts. But stochastic time-series values become **uncorrelated noise** at large lags - violating this premise.

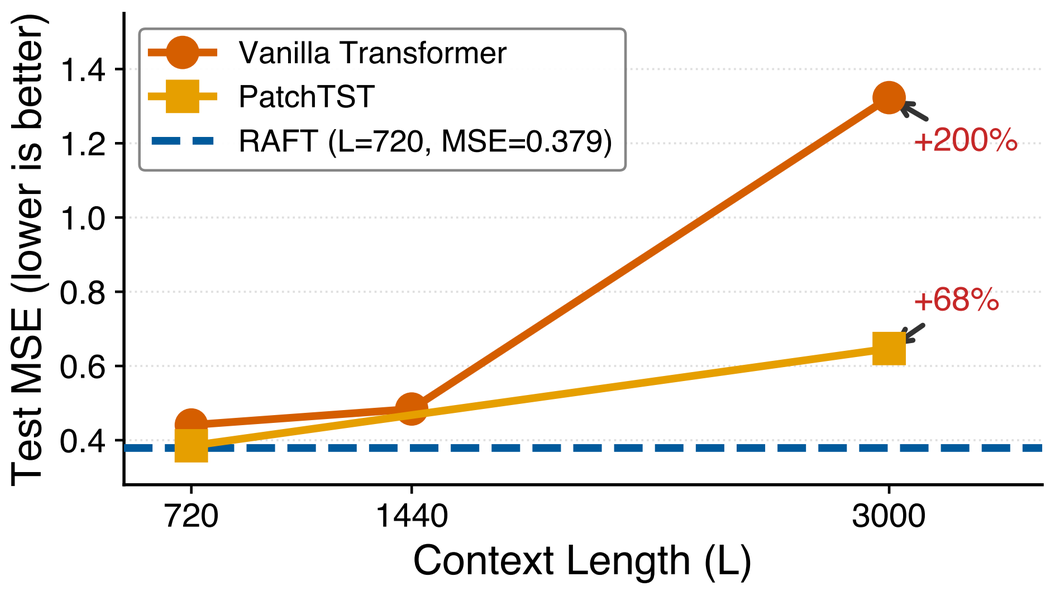
Does extending context length always improve time series forecasting accuracy?

We demonstrate:

- An **Inverse Scaling Law** - MSE *rises* as L grows
- Selective retrieval (RAFT, Han et al., 2025) outperforms window expansion
- Validated on **3 datasets** × **3 horizons**

The Inverse Scaling Law

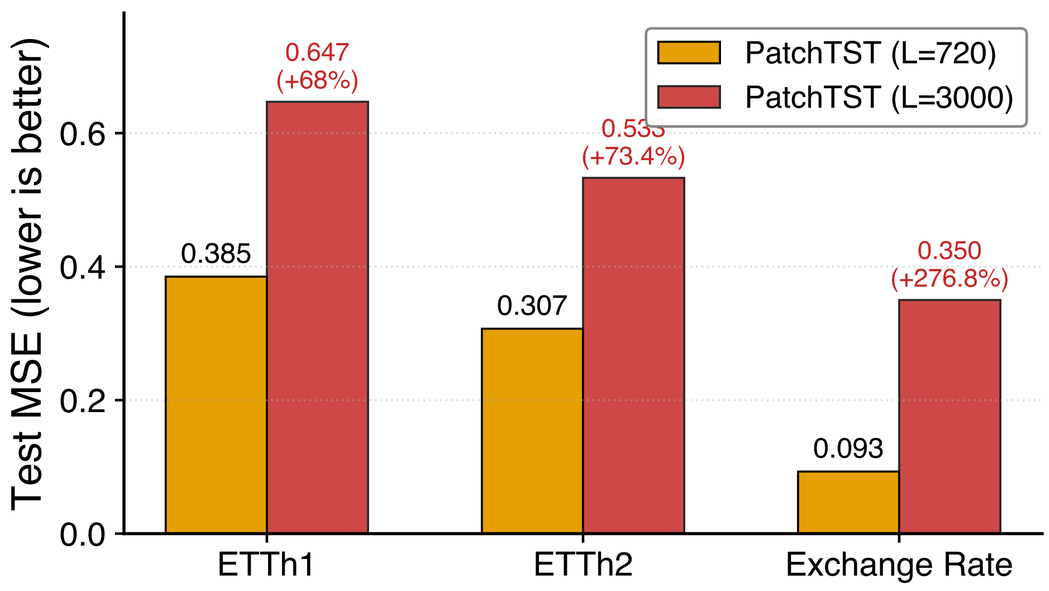
MSE rises monotonically with context length on ETTh1 (H=96):



PatchTST (Nie et al., 2023) degrades **+68%** (0.385 → 0.647)
Vanilla Transformer degrades **+200%** (0.441 → 1.323)
RAFT (Han et al., 2025) achieves **best MSE (0.379)** with fixed L=720

Cross-Domain Validation

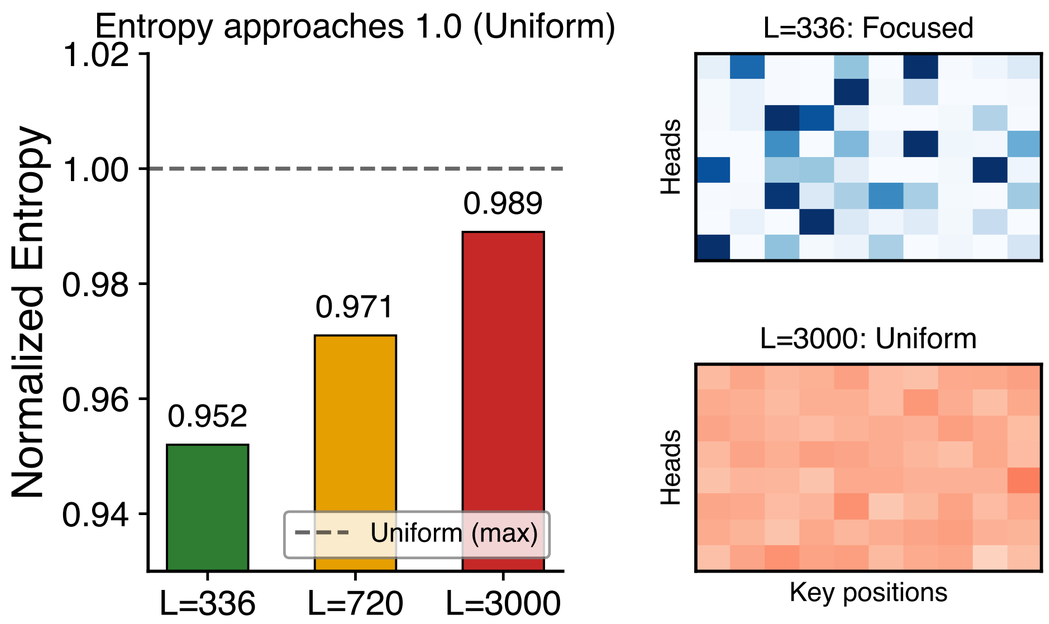
Long-context degradation is not a quirk of ETTh1. PatchTST (Nie et al., 2023) tested at L=720 vs. L=3000 across three domains:



Financial data (Exchange Rate) shows **+276.8%** degradation - high stochastic volatility amplifies noise accumulation.

Mechanism: Attention Entropy Collapse

Why does long context fail? The Softmax denominator grows with L, diluting probability mass across positions that carry no forecasting signal.



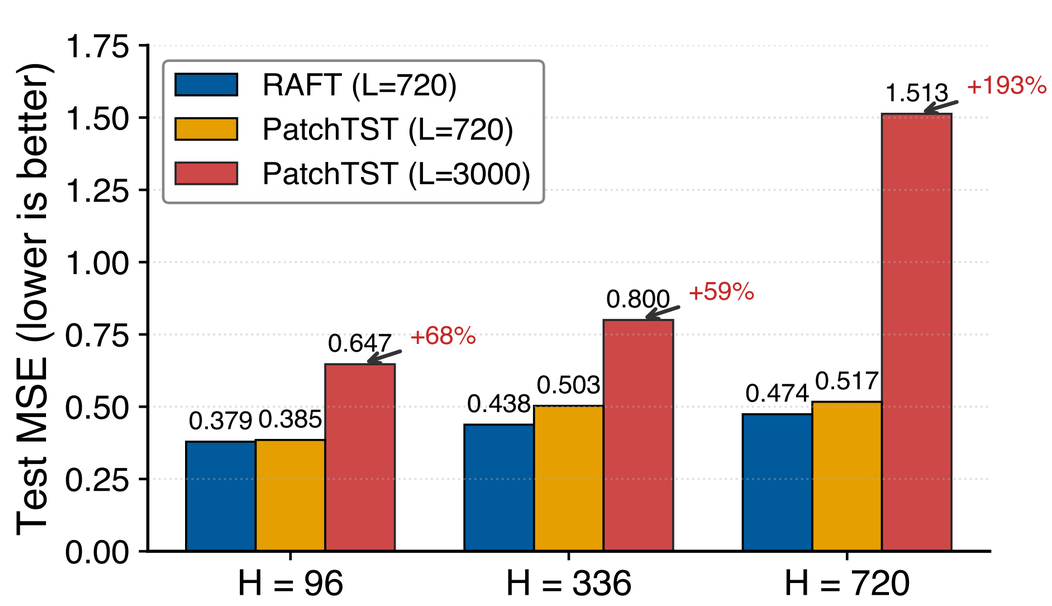
At L=3000 normalised entropy reaches **0.989** - within 1.1% of the theoretical max (1.0 = uniform distribution). RAFT's top-k retrieval acts as a **"hard attention" gate**, rejecting noise *before* the Softmax operates.

Resources & Contact

Code & Data | Contact | TSALM Workshop

Multi-Horizon Results

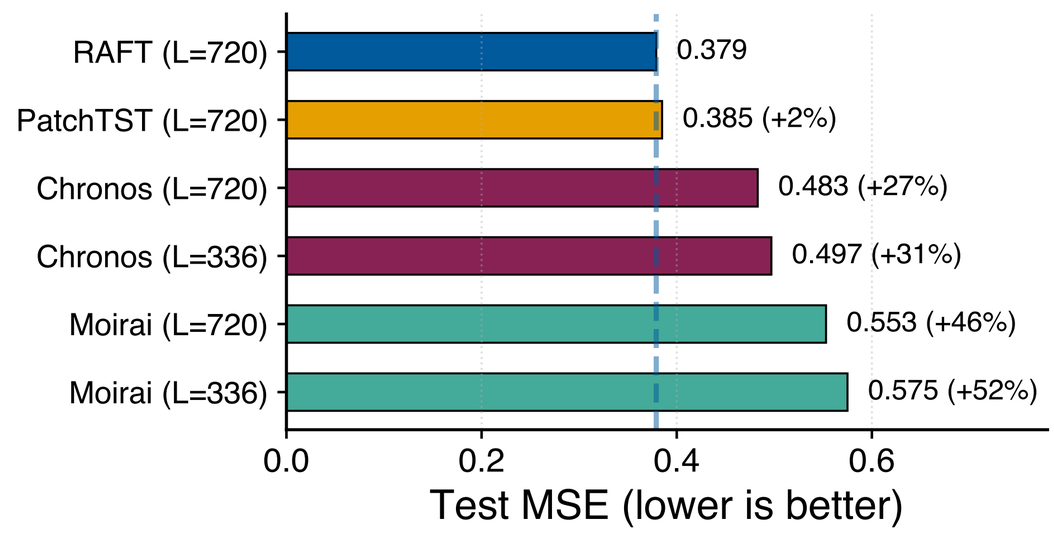
The inverse scaling law holds at every tested prediction horizon:



PatchTST degradation **worsens** with horizon: **+68%** (H=96), **+59%** (H=336), **+193%** (H=720). RAFT (Han et al., 2025) holds the lowest MSE everywhere.

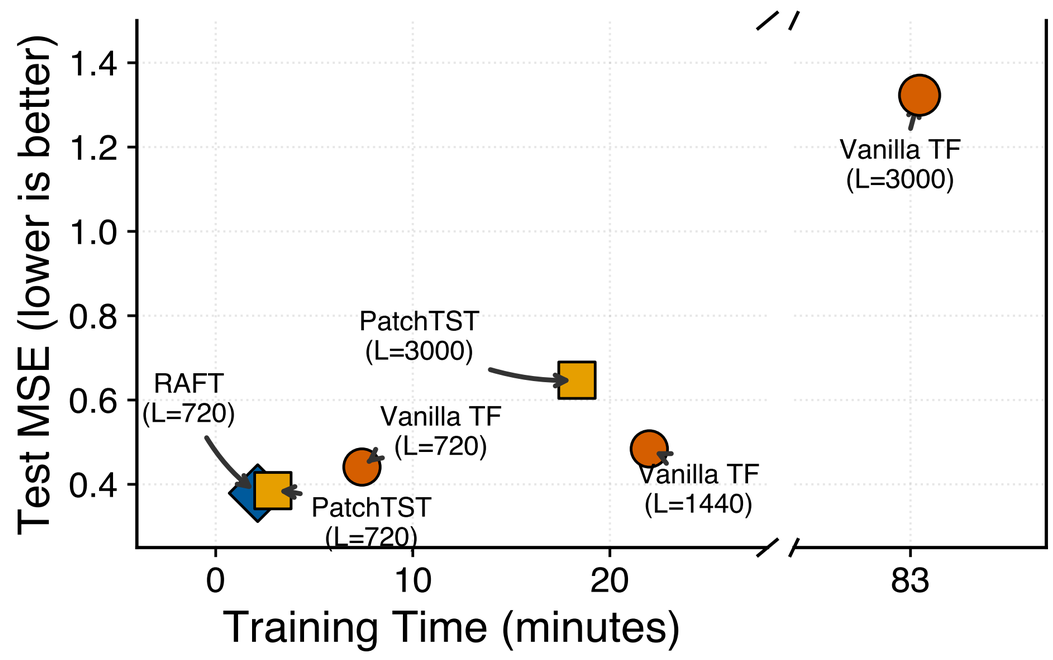
Foundation Model Comparison

Zero-shot foundation models also fall short of targeted retrieval:



Chronos-T5-Small (Ansari et al., 2024): MSE 0.483 **(+27%)**. Moirai-1.1-R-Small (Woo et al., 2024): MSE 0.553 **(+46%)**. Zero-shot generality does **not** compensate for selective retrieval.

Efficiency



RAFT: **2.13 min** / MSE **0.379**. Vanilla TF (L=3000): 83.47 min / MSE 1.323. **~40× faster, 71% lower error.**

Conclusion

Inverse Scaling Law: extending context inflates attention entropy and degrades accuracy.

Confirmed across 3 datasets, 3 horizons, and 3 model families.

Context quality > context volume.

Future work: Integrate retrieval into pretraining with dynamic retrieval heads.

References: [1] Han et al., "RAFT: Retrieval-Augmented Forecasting Transformer," 2025. [2] Nie et al., "A Time Series is Worth 64 Words: Long-term Forecasting with Transformers," NeurIPS, 2023. [3] Ansari et al., "Chronos: Learning the Language of Time Series," 2024. [4] Woo et al., "Unified Training of Universal Time Series Forecasting Transformers (Moirai)," 2024. [5] Zhou et al., "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting," AAAI, 2021.