



# Position: Time-Series Foundation Models Require Explicit Domain-Level Benchmarks

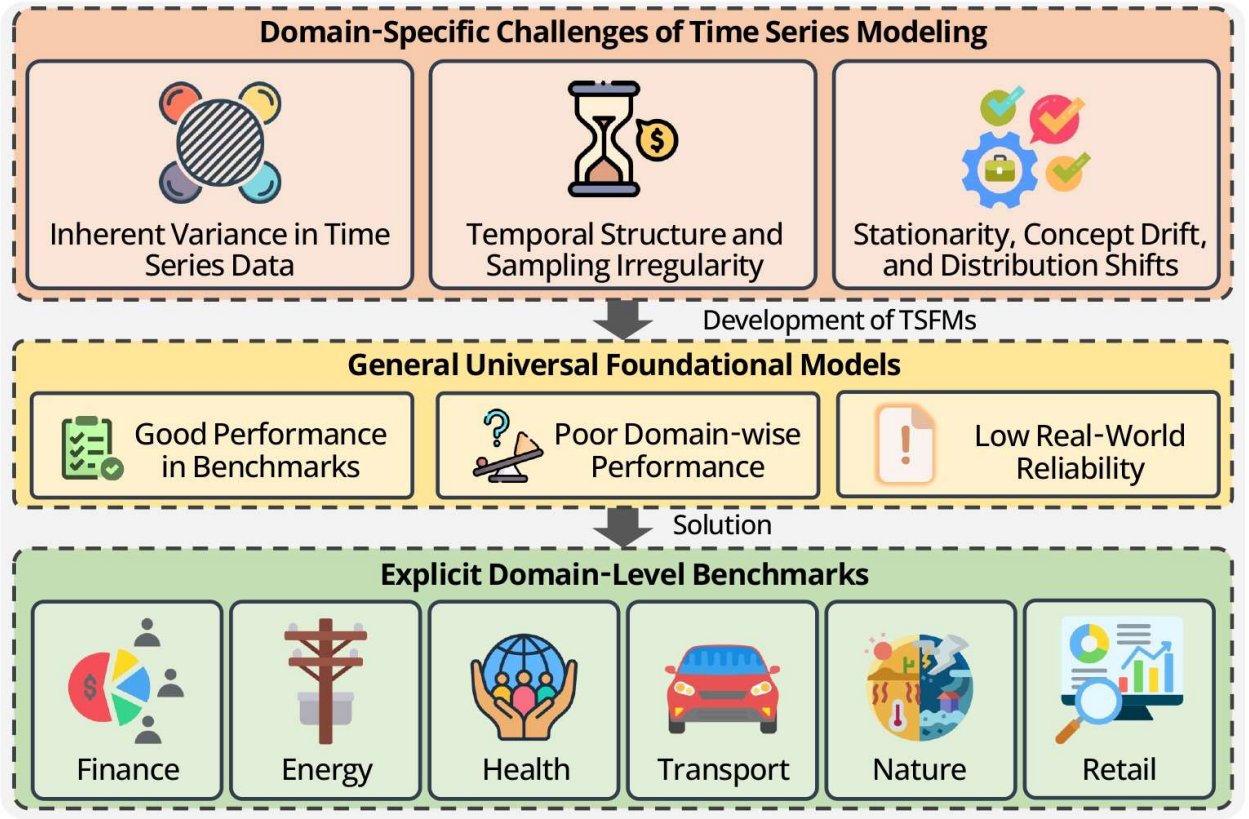
Md Asif Bin Syed

Md Younus Ahamed

Azmine Toushik Wasi

# Position

**Domain-specific differences in time series structure limit universal TSFMs, producing strong benchmark averages but inconsistent domain-wise reliability.** We therefore argue for explicit domain-level benchmarks to expose transfer failures and enable meaningful evaluation



# Domain-Specific Challenges of TSFMs

1

INHERENT VARIANCE IN TIME SERIES DATA

2

TEMPORAL STRUCTURE AND SAMPLING IRREGULARITY

3

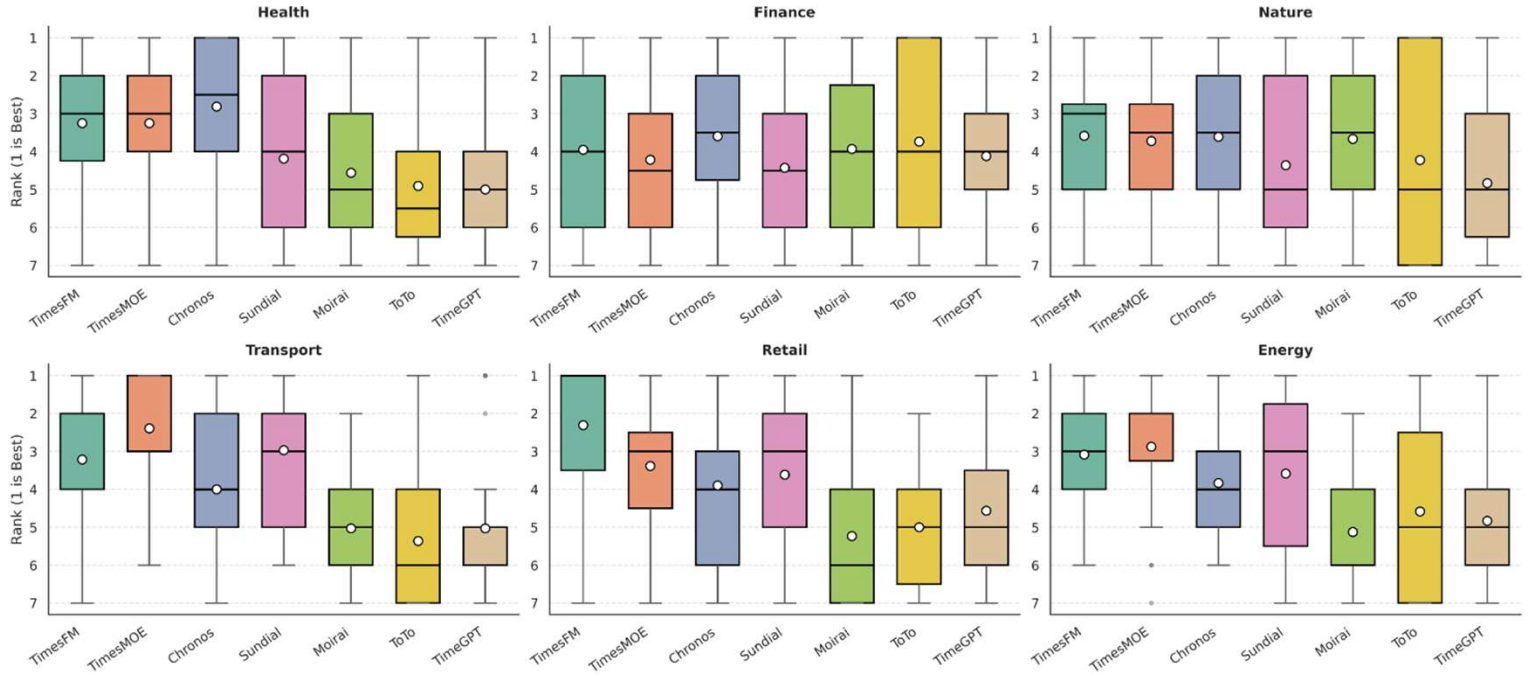
STATIONARITY, CONCEPT DRIFT, AND DISTRIBUTION SHIFTS

# Why Existing TSFM Benchmarks Fail?

**Existing TSFM benchmarks are highly domain-imbalanced, so aggregate scores can hide weak performance in underrepresented domains and motivate domain-level evaluation.**

| | GIFT-Eval | | TSFM-Bench | |
|---|---|---|---|---|
| **Domain** | **# Series (%)** | **# Obs (%)** | **# Series (%)** | **# Obs (%)** |
| Economic/Finance | 100.0K (69.3%) | 25.3M (16.0%) | 236 (4.9%) | 238.8K (0.6%) |
| Energy | 2.0K (1.4%) | 74.1M (46.9%) | 493 (10.3%) | 17.0M (42.2%) |
| Nature | 32.6K (22.6%) | 3.2M (2.0%) | 54 (1.1%) | 1.9M (4.8%) |
| Healthcare | 1.0K (0.7%) | 129.4K (0.1%) | 955 (20.0%) | 1.3M (3.3%) |
| Sales | 3.7K (2.6%) | 671.7K (0.4%) | – | – |
| Traffic/Transport | 1.3K (0.9%) | 38.0M (24.1%) | 1.0K (21.6%) | 18.2M (45.2%) |
| Web | 3.5K (2.4%) | 16.6M (10.5%) | 2.0K (41.9%) | 1.6M (3.9%) |
| **Total** | **144.2K** | **158.0M** | **4.8K** | **40.2M** |

# No TSFM Dominates all Domains




**The box plots summarize the MAE-based ranking of each TSFM within each domain, where a lower rank indicates better performance. Rank 1 represents the best-performing model.**

# Alternative view & Call to Action