



**BNP PARIBAS**

**CORPORATE & INSTITUTIONAL BANKING**

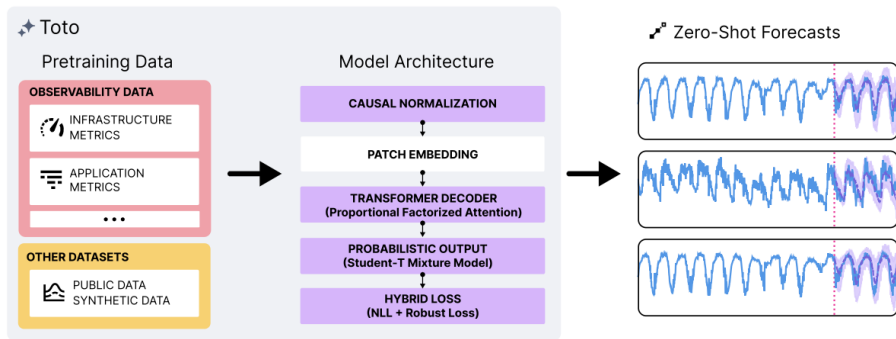
# This Time is Different: An Observability Perspective on Time Series Foundation Models

## Paper Review

MARZOUG AYOUB

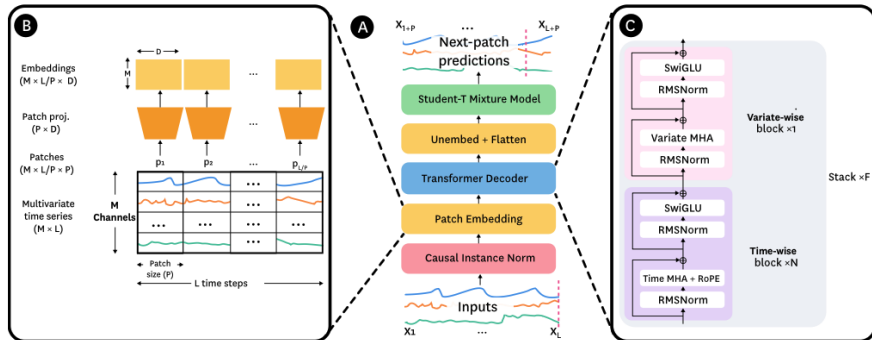
June 26th, 2025

# What is TOTO?



**TOTO** (**T**ime **S**eries **O**ptimized **T**ransformer for **O**bservability) : FM specifically designed for forecasting observability data, with 151M parameters.

# Backbone Architecture



# Patch-based Causal Scaling

## Problem

Standard LayerNorm is **non-causal**  $\Rightarrow$  information leakage in autoregressive settings.

## Solution

Scaling factors for each patch are computed exclusively from the current patch and past data :

$$\hat{\mu}_t = \frac{\sum_{i=1}^t w_i x_i}{\sum_{i=1}^t w_i}, \quad \hat{s}_t = \sqrt{\frac{\sum_{i=1}^t w_i (x_i - \hat{\mu}_t)^2}{\sum_{i=1}^t w_i - 1}} + 0.1$$

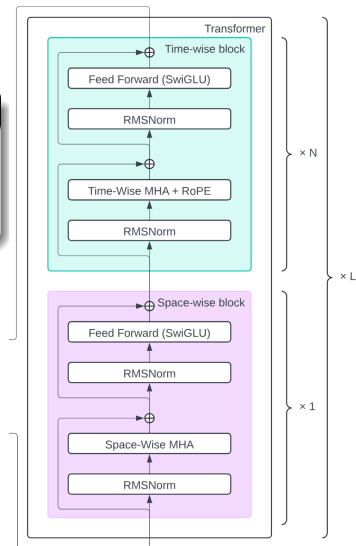
# Proportional Factorized Space-Time Attention

## Motivation

Standard attention over flattened  $[T, D]$  sequences is inefficient for high-dimensional multivariate time series

## Factorized Attention alternates :

- **Time-wise attention** over patches across time.
- **Variate-wise attention** over variables across channels.



## Forecasting Mechanism

Generating **probabilistic forecasts** via a **Student-T Mixture Model (SMM)** head :

$$p(x) = \sum_{k=1}^K \pi_k \mathcal{T}(x \mid \mu_k, \tau_k, \nu_k)$$

More robust than Gaussian or quantized token prediction — captures heavy tails and outliers in observability data.

# Pre-training Loss

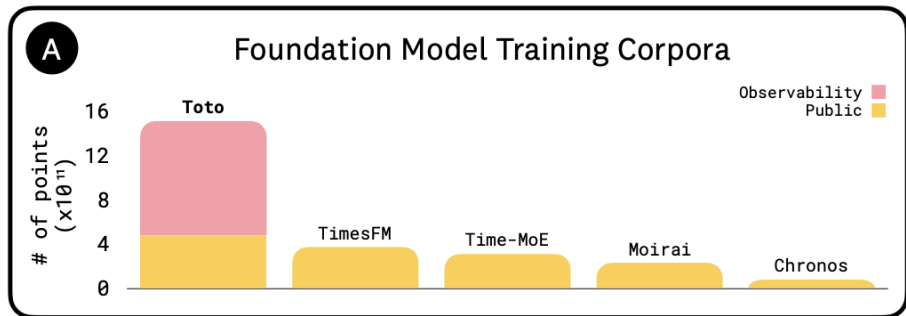
## Composite Robust loss

Combining SMM NLL and robust Cauchy loss :

$$\mathcal{L} = \lambda \cdot \mathcal{L}_{\text{NLL}} + (1 - \lambda) \cdot \log \left( 1 + \frac{(x - \hat{x}_t)^2}{2\delta^2} \right)$$

where  $\lambda \in [0, 1]$  is a ratio tuned purely for autoregressive forecasting, with optimal value  $\lambda = 0.57$ .

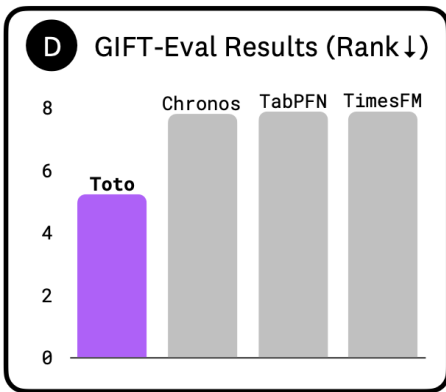
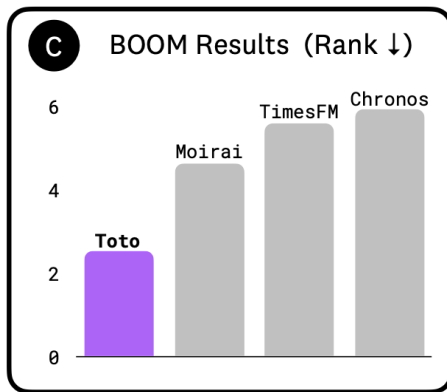
# Pretraining Corpora



*Largest and most diverse training corpus (2.36 trillion points).*



# Boom Benchmark & Experiments



*TOTO outperforms state-of-the-art models by  $\sim 12\%$  CRPS.*

# Potential for financial time series

- **Covariates Handling** : Natively supports multivariate input — ideal for integrating multiple financial indicators and correlated instruments.
- **Adaptable flexible loss** : can easily encode finance-specific priors (e.g. OU mean-reversion or GARCH volatility) :

$$\mathcal{L} = \lambda_1 \underbrace{\mathcal{L}_{\text{NLL}}}_{\text{Student-T mixture}} + \lambda_2 \underbrace{\mathcal{L}_{\text{Cauchy}}}_{\text{robust point error}} + \lambda_3 \underbrace{\mathcal{L}_{\text{prior}}}_{\text{OU / GARCH / ...}}$$

- **Frequency Agnostic** : Patch-based encoder handles high-frequency streams naturally ;
  - *Low-freq* : larger patch size and/or add seasonal event covariates.
  - *High-freq* : down-sample slightly.

# Questions ?

*Thank you for your attention !*