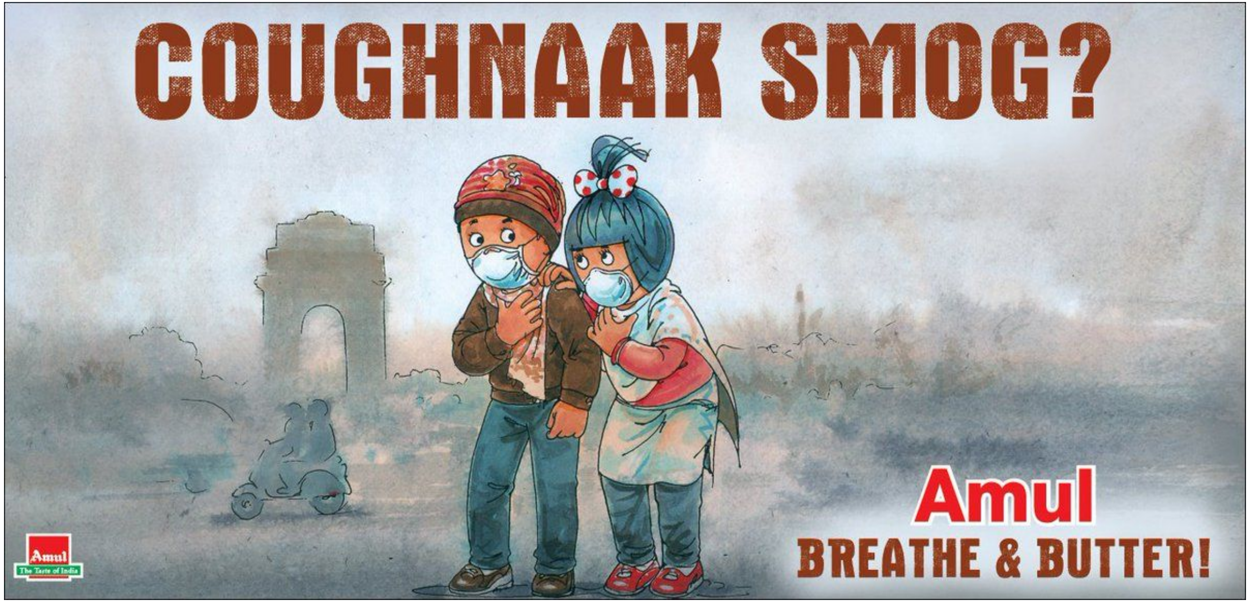
Explo(ring)iting Foundational Time Series Models for Data Forecasting

Name : **Harshit** (MSR)
Entry Number : **2024SIY7587**

Background

Time-series foundation models claim strong zero shot forecasting performance but understanding how they behave under real-world computational constraints.

Forecasting TS Models on Weather Dataset, across Pollution, and other covariates.



NeurIPS'24

Are Language Models Actually Useful for Time Series Forecasting?

Mingtian Tan
University of Virginia
wtd3gz@virginia.edu

Mike A. Merrill
University of Washington
mikeam@cs.washington.edu

Vinayak Gupta
University of Washington
vinayak@cs.washington.edu

Tim Althoff
University of Washington
althoff@cs.washington

Thomas Hartvigsen
University of Virginia
hartvigsen@virginia.edu

Data

Vayu

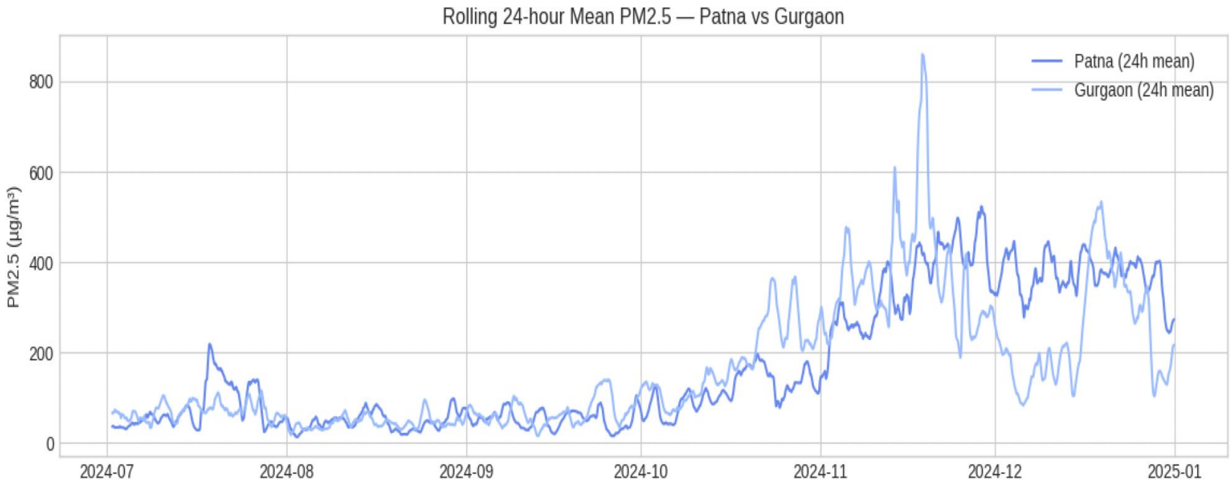
- Vayu is an initiative by UNDP, under Open Digital Stack on Air Pollution for hyperlocal mapping of air pollution.
- Datasets of two cities :
  - Patna
  - Gurgaon

Keywords

- Context : Context Length Days {2,4,6,8,12}
- Horizon : Forecast Horizon Hours
- Latency : (per ms)
- Throughput : (Req/Sec) : GFLOPS/sec
- Strides

Dataset Details

:



$PM_{2.5}$ concentration as primary target variable

Hourly calibrated sensor readings, includes meteorological other covariates

Chronos (TMLR'24)

Published in Transactions on Machine Learning Research (10/2024)

Chronos: Learning the Language of Time Series

Abdul Fatir Ansari[1,*], Lorenzo Stella[1,*], Caner Turkmen[1], Xiyuan Zhang[3,†], Pedro Mercado[1], Huibin Shen[1], Oleksandr Shchur[1], Syama Sundar Rangapuram[1], Sebastian Pineda Arango[4,†], Shubham Kapoor[1], Jasper Zschiegner[†], Danielle C. Maddix[1], Hao Wang[1,5,†], Michael W. Mahoney[2,6,†], Kari Torkkola[2], Andrew Gordon Wilson[2,7,†], Michael Bohlke-Schneider[1], Yuyang Wang[1]

{ansarnd, stellalo}@amazon.com

[1]AWS AI Labs, [2]Amazon Supply Chain Optimization Technologies, [3]UC San Diego, [4]University of Freiburg, [5]Rutgers University, [6]UC Berkeley, [7]New York University

Models used :

Model	Params	Base T5	Typical strengths
chronos-t5-tiny	~8M	t5-efficient-tiny	very low latency and footprint
chronos-t5-small	~46M	t5-efficient-small	good middle ground
chronos-t5-base	~200M	t5-efficient-base	highest capacity and generalization

#Models : 3 (varying on their size offering, each one is build on top of T5.

Research Questions

1. Zero Shot Performance Generalization Capacity under ICL.
2. Context Length, Horizon variations on Accuracy.
3. Scaling Behaviour : Quality-Compute Correlation (Test time scaling ?)

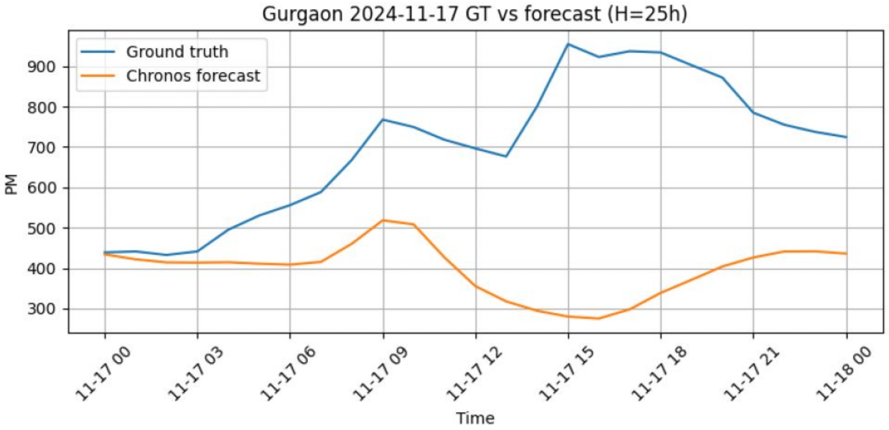
Accuracy Evaluation :

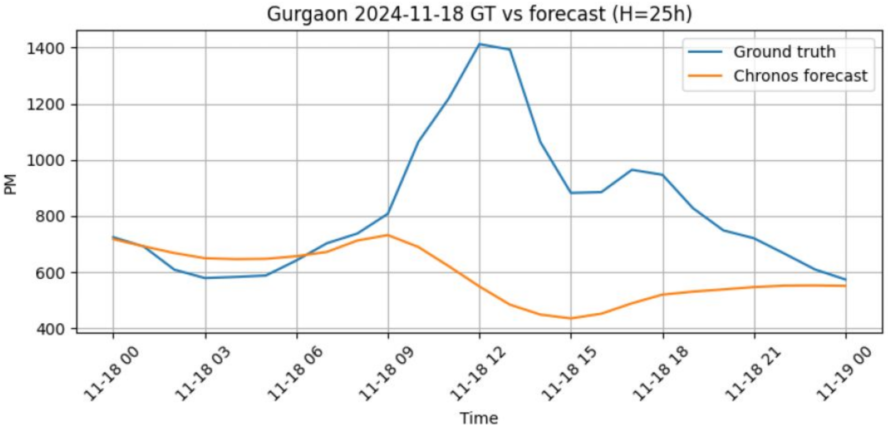
- Root Mean Squared Error (L2 Norm)
- MAE (L1)
-

System Profiling :

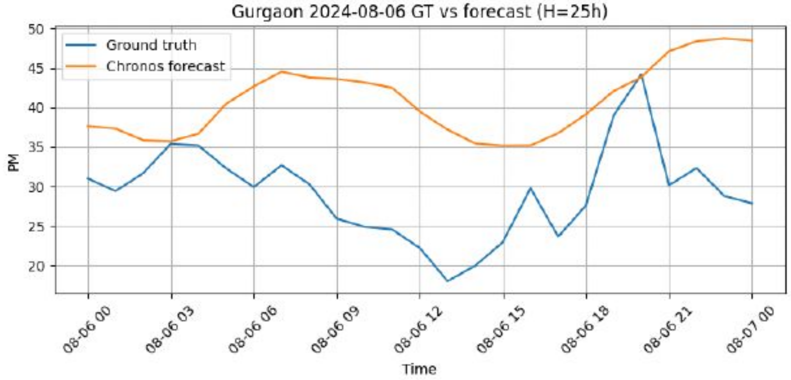
- Latency
- Flops : Floating Point Operations.
- Gflops : Giga Flops.
- RSS: Residence Set Size (RAM usage/working memory)
- IPC
- Cache Miss Rates
- Branch Miss Rates

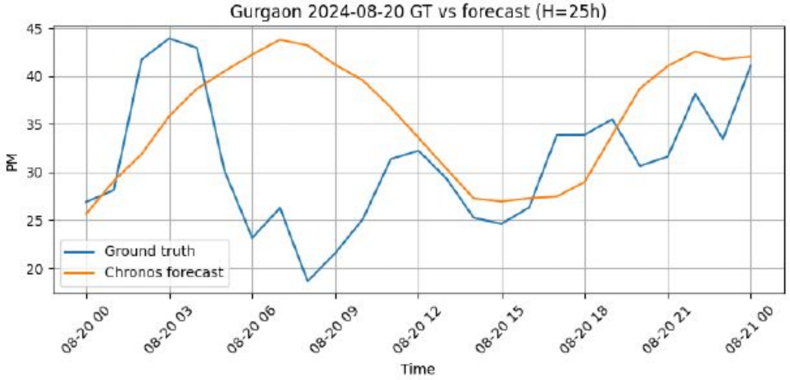
Max PM


Gurgaon 2024-11-17 GT vs forecast (H=25h)
Ground truth
Chronos forecast
PM
Time
900
800
700
600
500
400
300
11-17 00
11-17 03
11-17 06
11-17 09
11-17 12
11-17 15
11-17 18
11-17 21
11-18 00


Gurgaon 2024-11-18 GT vs forecast (H=25h)
Ground truth
Chronos forecast
PM
Time
1400
1200
1000
800
600
400
11-18 00
11-18 03
11-18 06
11-18 09
11-18 12
11-18 15
11-18 18
11-18 21
11-19 00

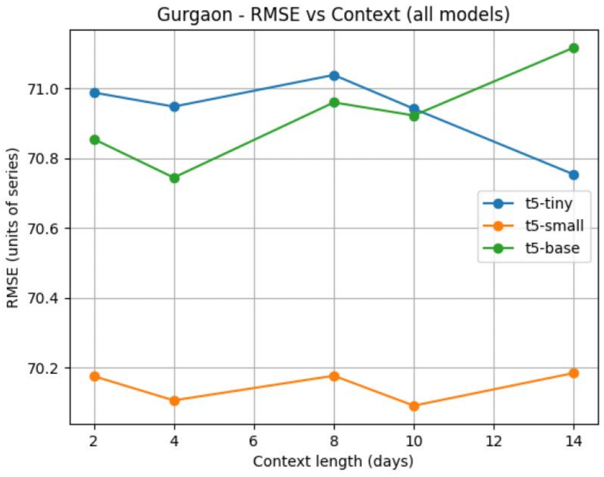
Min PM

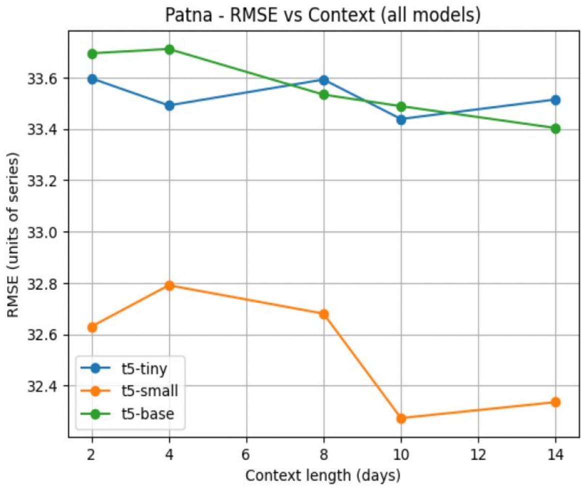

Gurgaon 2024-08-06 GT vs forecast (H=25h)
Ground truth
Chronos forecast
PM
Time
50
45
40
35
30
25
20
08-06 00
08-06 03
08-06 06
08-06 09
08-06 12
08-06 15
08-06 18
08-06 21
08-07 00


Gurgaon 2024-08-20 GT vs forecast (H=25h)
Ground truth
Chronos forecast
PM
Time
45
40
35
30
25
20
08-20 00
08-20 03
08-20 06
08-20 09
08-20 12
08-20 15
08-20 18
08-20 21
08-21 00

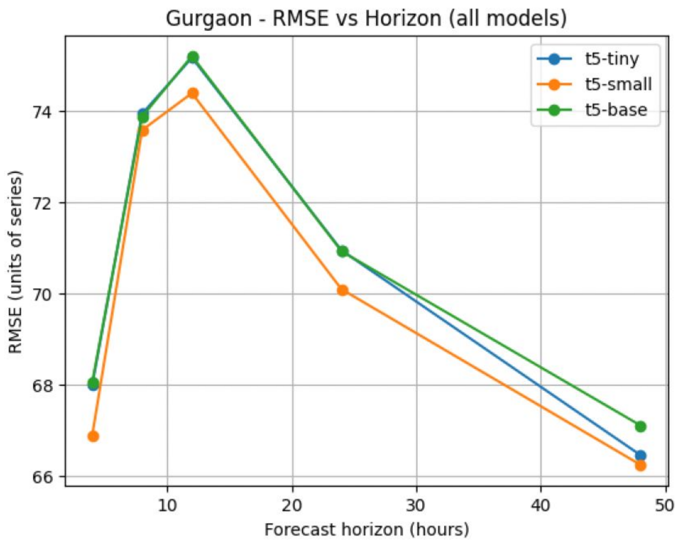
Experiment 1

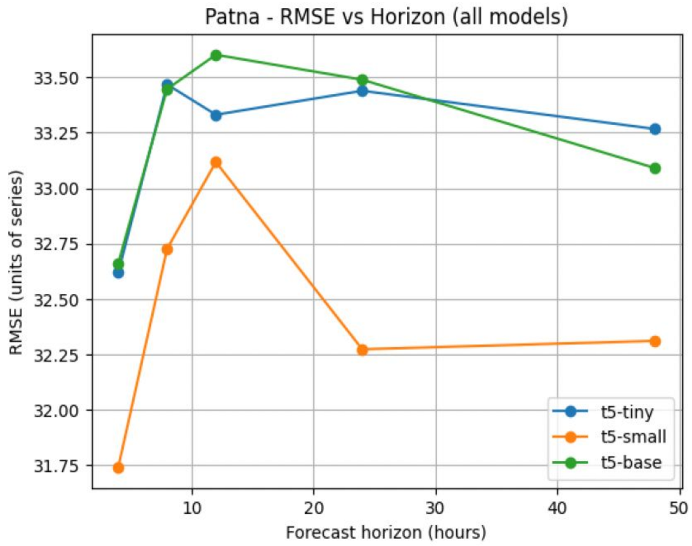
Assessing accuracy across different fixed context lengths and horizons to determine the optimal setting.





RMSE vs Forecast Horizon





Experiment 2 : Instance Profiling

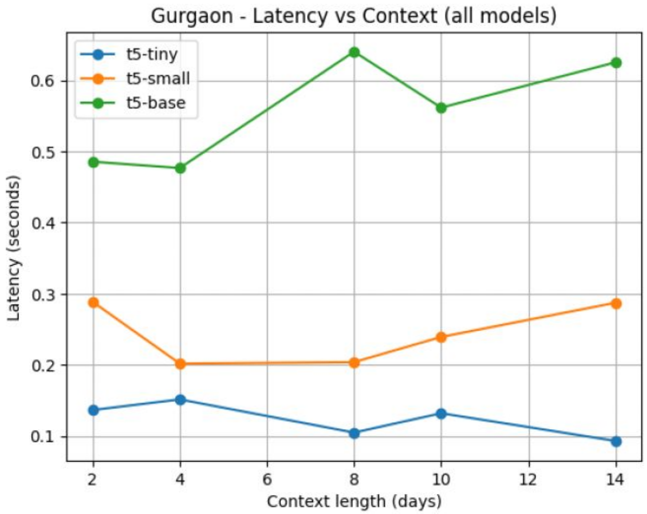
How was Profiling done?

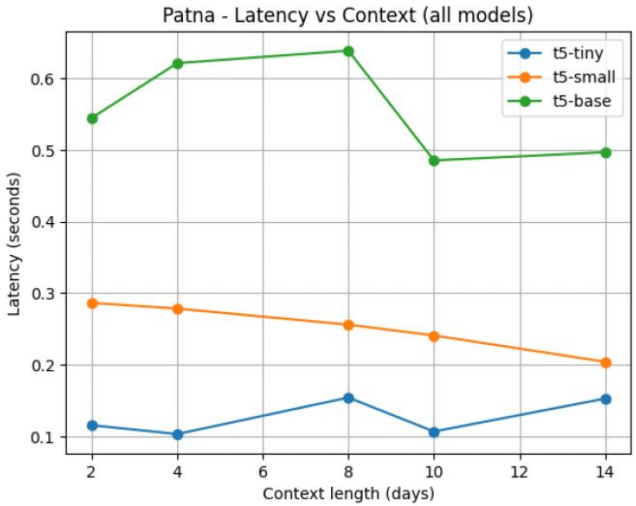
- Pytorch perf tool(Profiler).
- Linux perf tool counters

PyTorch Profiler

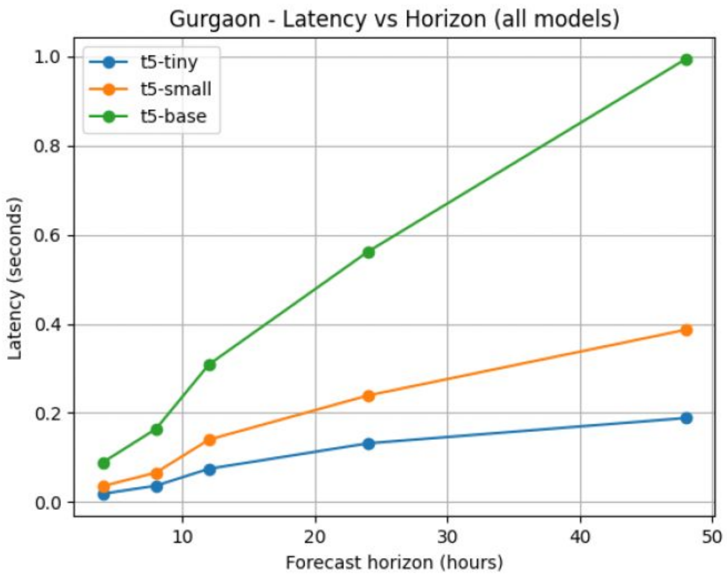
Latency vs Context Length

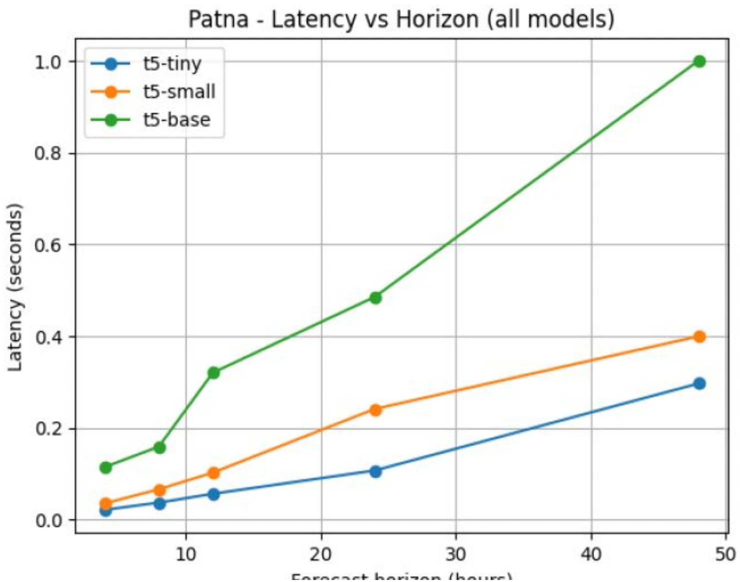
- Evaluating Latency, RSS, IPC, Cache Miss Rate, Branch Miss Rate, Cycle vs Context Length.



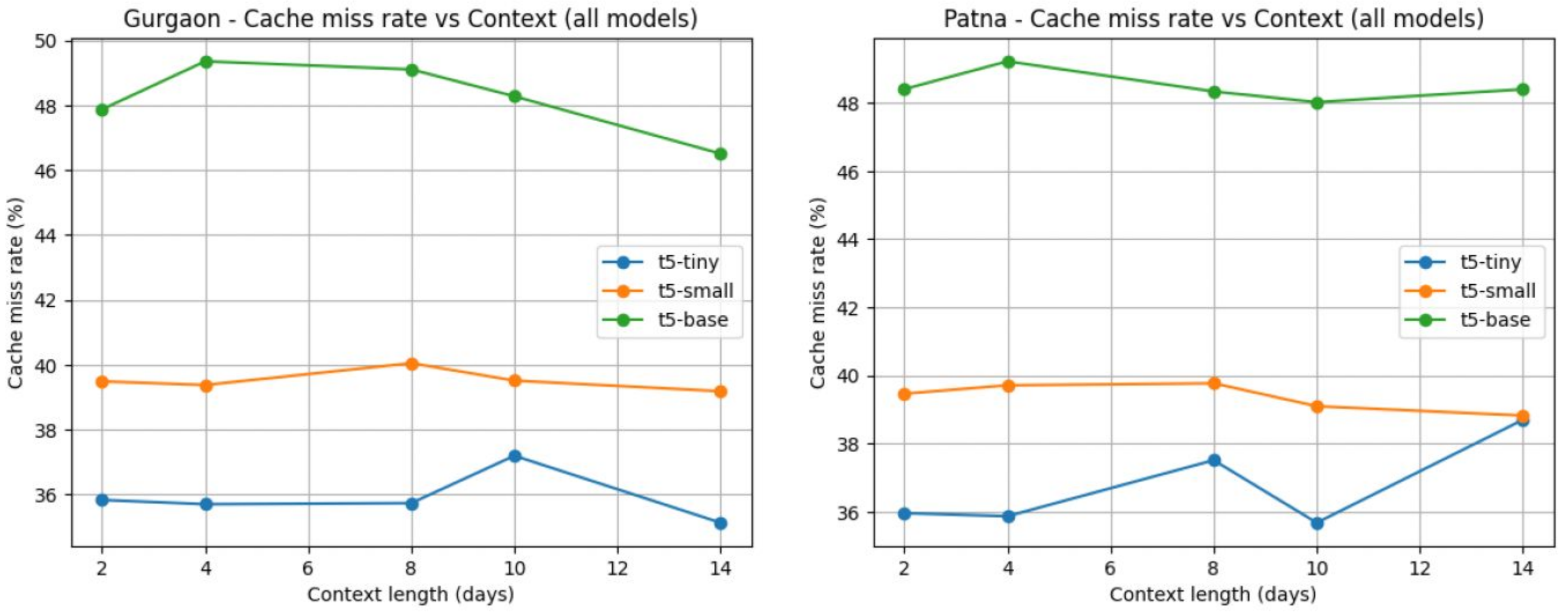


Latency vs Horizon (10 day Context)





Cache Miss Rates vs Context



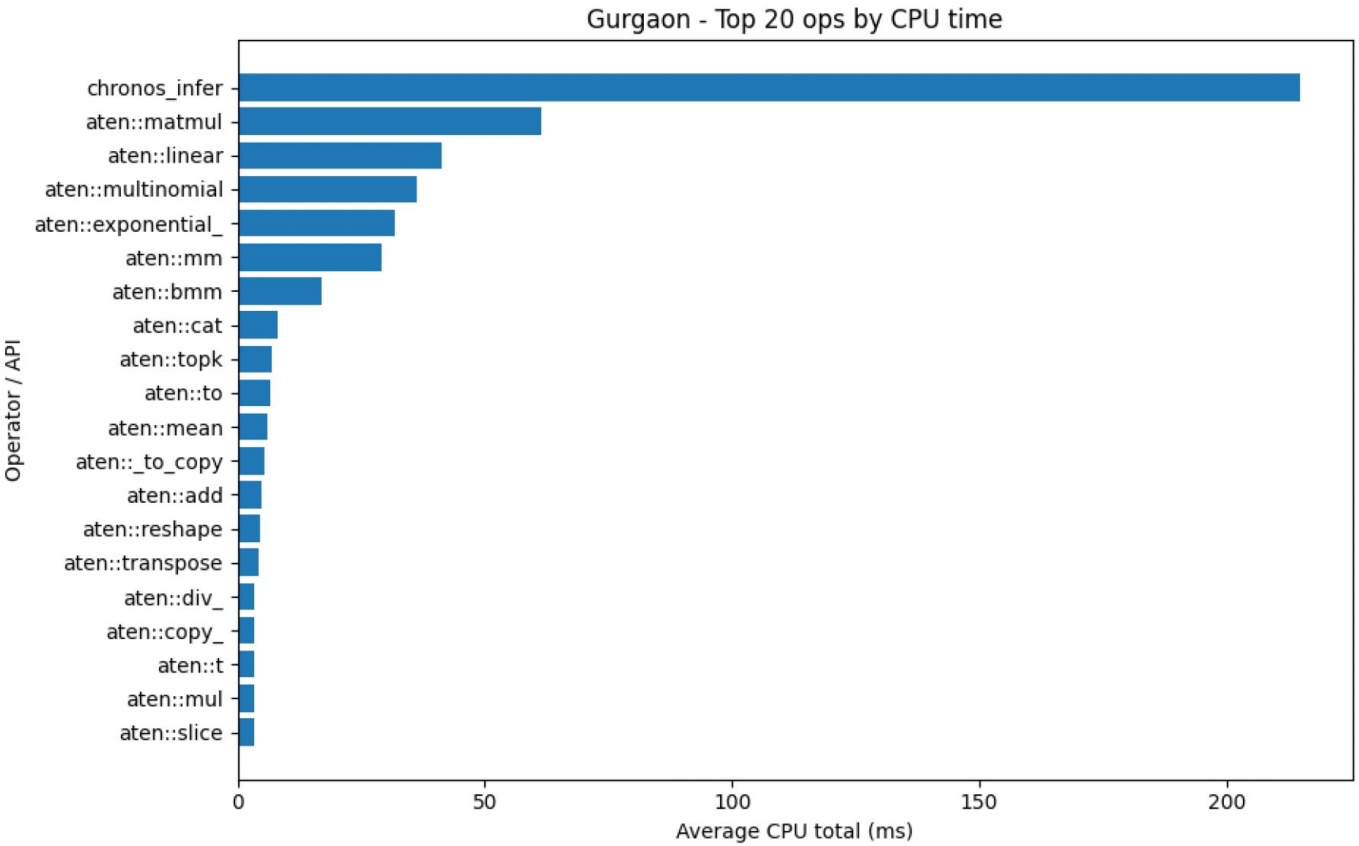
(a) Gurgaon (b) Patna

Experiment 3 : Operator Level Breakdown

Attention: 40-50% of CPU time across all models

Rest:

- Linear projections: 25-30% (Q/K/V transformations)
- LayerNorm: 8-12% (increases with depth)
- Activation functions: 5-8% (GELU/ReLU)


Gurgaon - Top 20 ops by CPU time
chronos_infer
aten::matmul
aten::linear
aten::multinomial
aten::exponential_
aten::mm
aten::bmm
aten::cat
aten::topk
aten::to
aten::mean
aten::_to_copy
aten::add
aten::reshape
aten::transpose
aten::div_
aten::copy_
aten::t
aten::mul
aten::slice
Operator / API
0
50
100
150
200
Average CPU total (ms)

Limitations and Future Work :

- Single Covariate.
- Chronos 2 or classifier like Xgboost.
- Generalization to other CPU and GPU architectures.

Closing Marks

Thank you for listening.
Any Questions?

Insights