Vertical AI Meets Observability

Ameet Talwalkar
12/7/25

Carnegie Mellon University

DATADOG

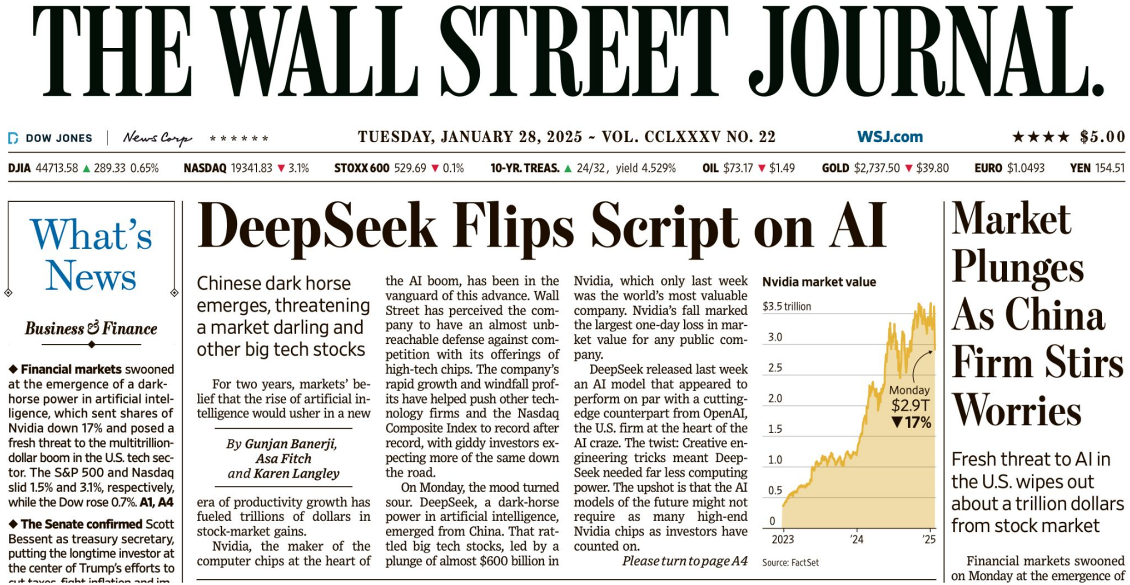
AI has taken over the world

Frontpage news

Startup valuations

Big tech spending

AI research pace



THE WALL STREET JOURNAL.

DOW JONES | News Corp ****** TUESDAY, JANUARY 28, 2025 ~ VOL. CCLXXXV NO. 22 WSJ.com ★★★★ $5.00

DJIA 44713.58 ▲ 289.33 0.65% NASDAQ 19341.83 ▼ 3.1% STOXX 600 529.69 ▼ 0.1% 10-YR. TREAS. ▲ 24/32, yield 4.529% OIL $73.17 ▼ $1.49 GOLD $2,737.50 ▼ $39.80 EURO $1.0493 YEN 154.51

What's News

Business & Finance

◆ **Financial markets** swooned at the emergence of a dark-horse power in artificial intelligence, which sent shares of Nvidia down 17% and posed a fresh threat to the multitrillion-dollar boom in the U.S. tech sector. The S&P 500 and Nasdaq slid 1.5% and 3.1%, respectively, while the Dow rose 0.7%. **A1, A4**

◆ **The Senate confirmed** Scott Bessent as treasury secretary, putting the longtime investor at the center of Trump's efforts to cut taxes, fight inflation and im-

DeepSeek Flips Script on AI

Chinese dark horse emerges, threatening a market darling and other big tech stocks

By Gunjan Banerji, Asa Fitch and Karen Langley

For two years, markets' belief that the rise of artificial intelligence would usher in a new era of productivity growth has fueled trillions of dollars in stock-market gains.

Nvidia, the maker of the computer chips at the heart of the AI boom, has been in the vanguard of this advance. Wall Street has perceived the company to have an almost unbreachable defense against competition with its offerings of high-tech chips. The company's rapid growth and windfall profits have helped push other technology firms and the Nasdaq Composite Index to record after record, with giddy investors expecting more of the same down the road.

On Monday, the mood turned sour. DeepSeek, a dark-horse power in artificial intelligence, emerged from China. That rattled big tech stocks, led by a plunge of almost $600 billion in Nvidia, which only last week was the world's most valuable company. Nvidia's fall marked the largest one-day loss in market value for any public company.

DeepSeek released last week an AI model that appeared to perform on par with a cutting-edge counterpart from OpenAI, the U.S. firm at the heart of the AI craze. The twist: Creative engineering tricks meant DeepSeek needed far less computing power. The upshot is that the AI models of the future might not require as many high-end Nvidia chips as investors have counted on.

Please turn to page A4

Nvidia market value

Source: FactSet

Market Plunges As China Firm Stirs Worries

Fresh threat to AI in the U.S. wipes out about a trillion dollars from stock market

Financial markets swooned on Monday at the emergence of

AI has taken over the world?

Hype

Credit: GPT-4o

'AI eating the world'

!=

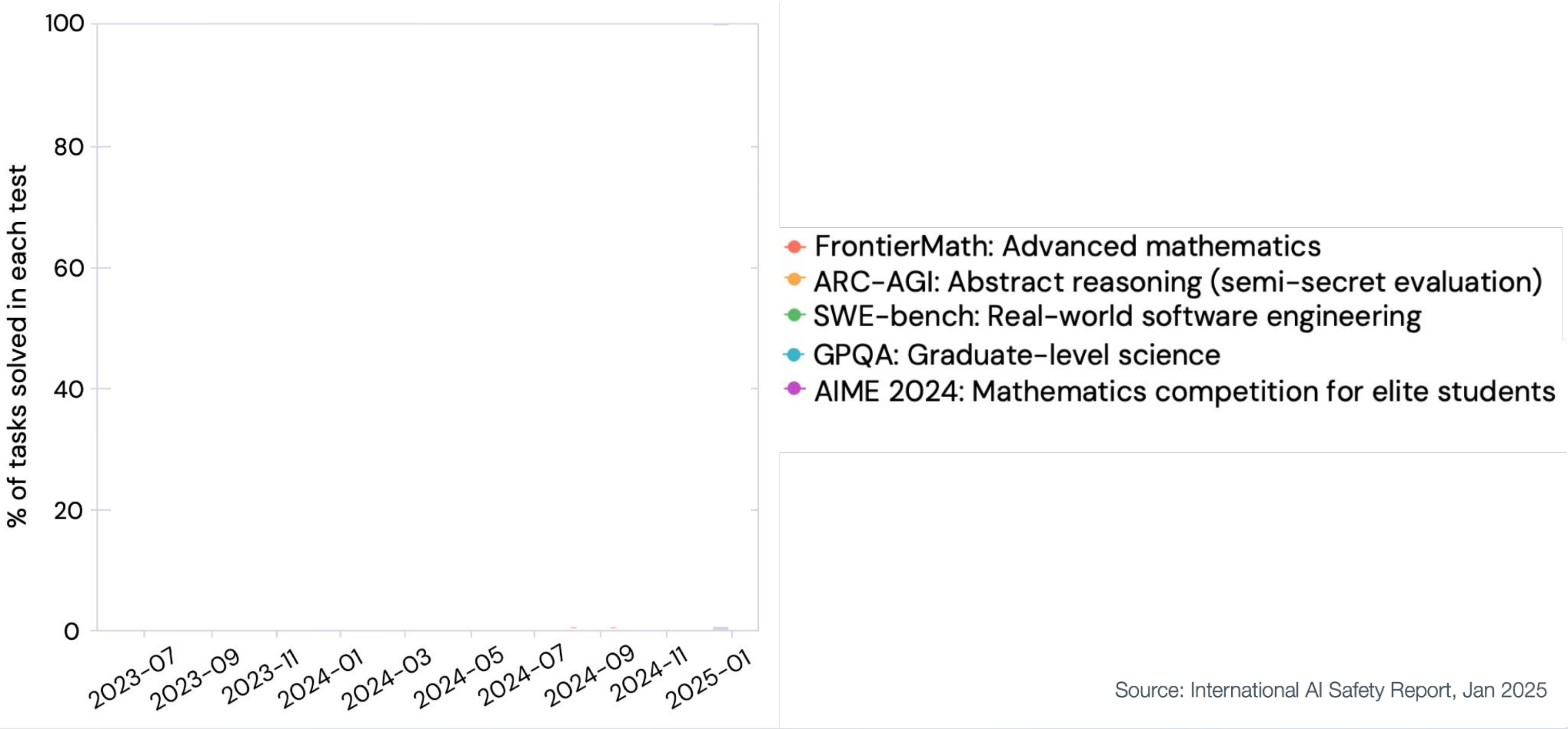
Impact

Search

Coding

Why?

Hypothesis 1: Just wait, it’s coming

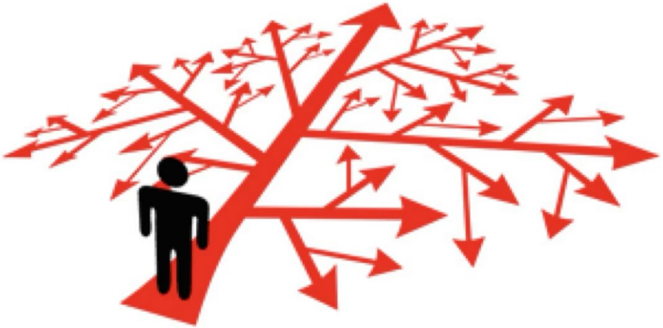


Source: International AI Safety Report, Jan 2025

Hypothesis 2: Specialization matters!

People specialize to become experts

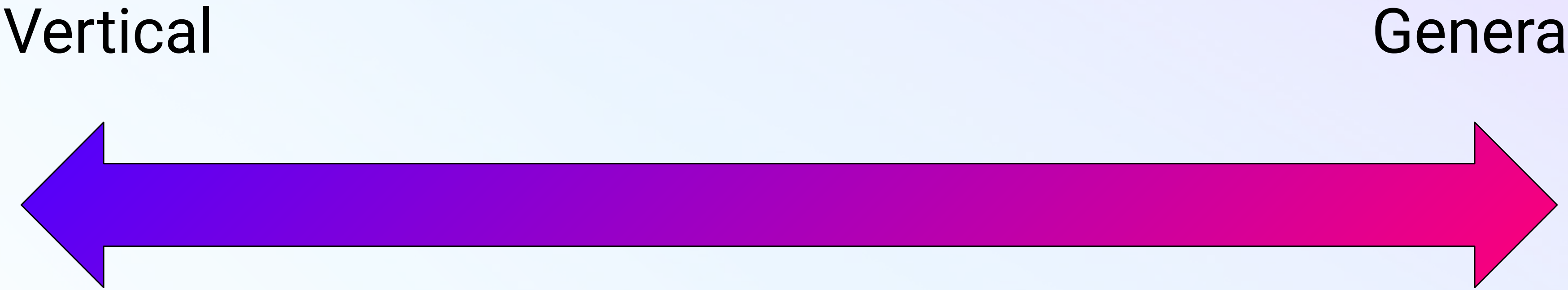
- E.g., scientists and athletes
- E.g., the human brain itself



AI will need to specialize

- Accuracy
- Efficiency

It’s not either/or

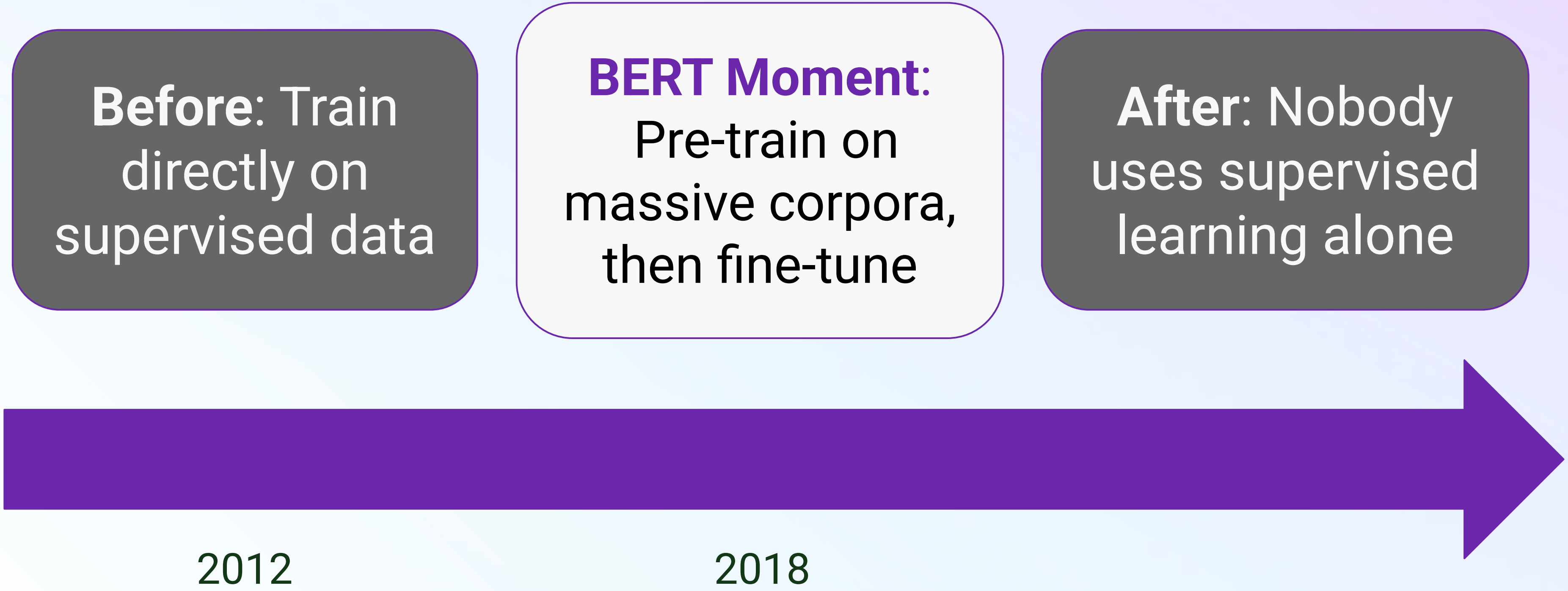


Vertical

General

How good are today’s specialized FMs?

Toto: An Observability TSFM



Before: Train directly on supervised data

BERT Moment: Pre-train on massive corpora, then fine-tune

After: Nobody uses supervised learning alone

2012

2018

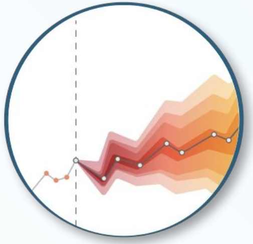
NLP Timeline

How good are today's Specialized FMs?

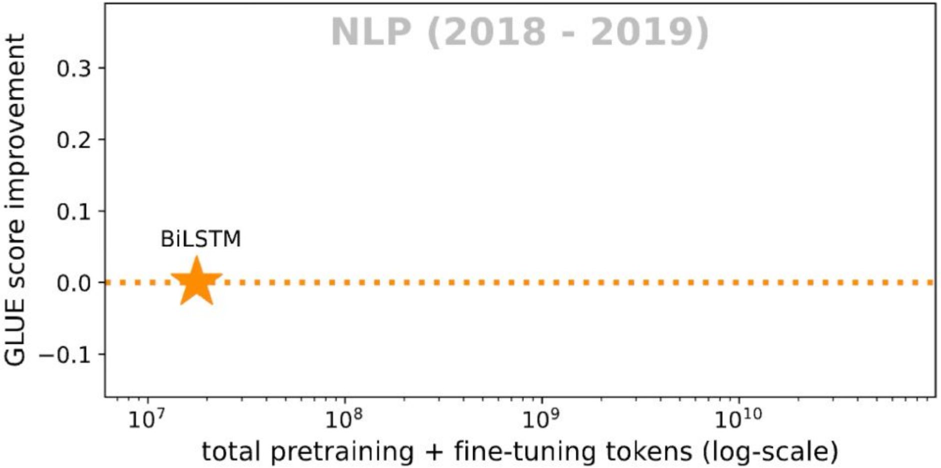
[Gupta*-Xu*-Cheng-Shen-Shen-**T**-Khodak, ICLR25]

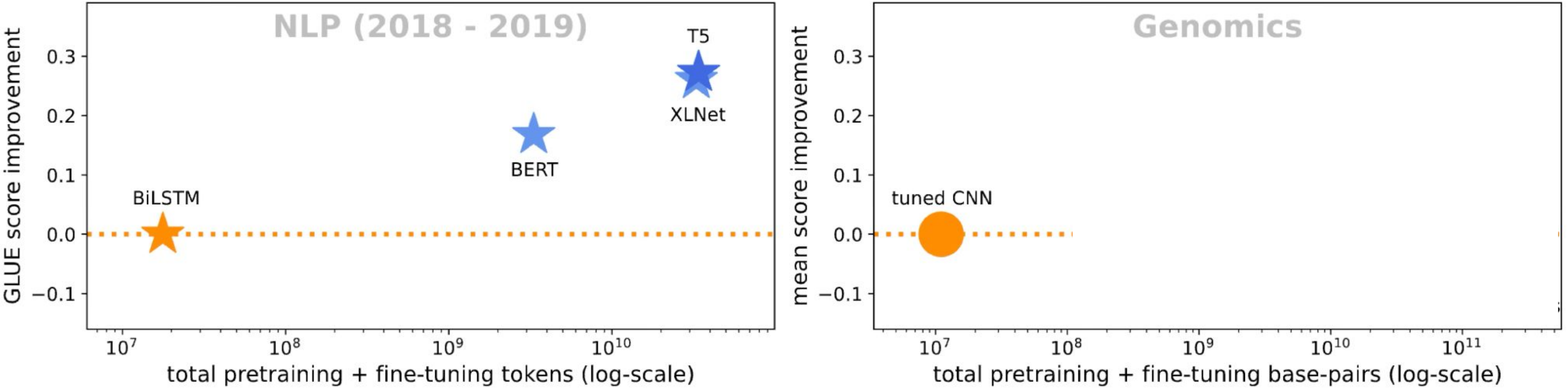
Genomics

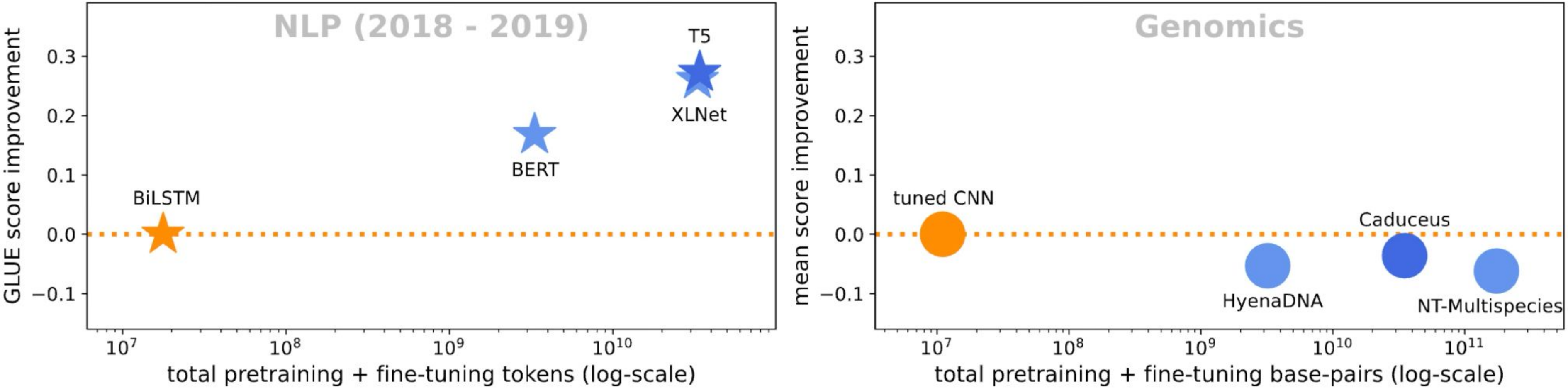
Satellite Imaging



Time Series


NLP (2018 - 2019)
GLUE score improvement
0.3
0.2
0.1
0.0
−0.1
BiLSTM
10^7
10^8
10^9
10^{10}
total pretraining + fine-tuning tokens (log-scale)


NLP (2018 - 2019)
GLUE score improvement
0.3
0.2
0.1
0.0
−0.1
T5
XLNet
BERT
BiLSTM
10^7
10^8
10^9
10^{10}
total pretraining + fine-tuning tokens (log-scale)
Genomics
mean score improvement
0.3
0.2
0.1
0.0
−0.1
tuned CNN
10^7
10^8
10^9
10^{10}
10^{11}
total pretraining + fine-tuning base-pairs (log-scale)

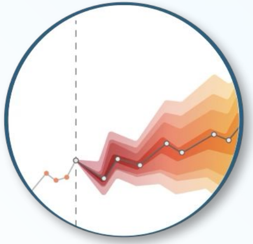

NLP (2018 - 2019)
GLUE score improvement
0.3
0.2
0.1
0.0
−0.1
T5
XLNet
BERT
BiLSTM
10^7
10^8
10^9
10^{10}
total pretraining + fine-tuning tokens (log-scale)
Genomics
mean score improvement
0.3
0.2
0.1
0.0
−0.1
tuned CNN
Caduceus
HyenaDNA
NT-Multispecies
10^7
10^8
10^9
10^{10}
10^{11}
total pretraining + fine-tuning base-pairs (log-scale)

How good are today's Specialized FMs?

[Gupta*-Xu*-Cheng-Shen-Shen-**T**-Khodak, ICLR25]

Genomics

Satellite Imaging



Time Series

Don't assume specialized FMs work!

Baselines & benchmarks are important

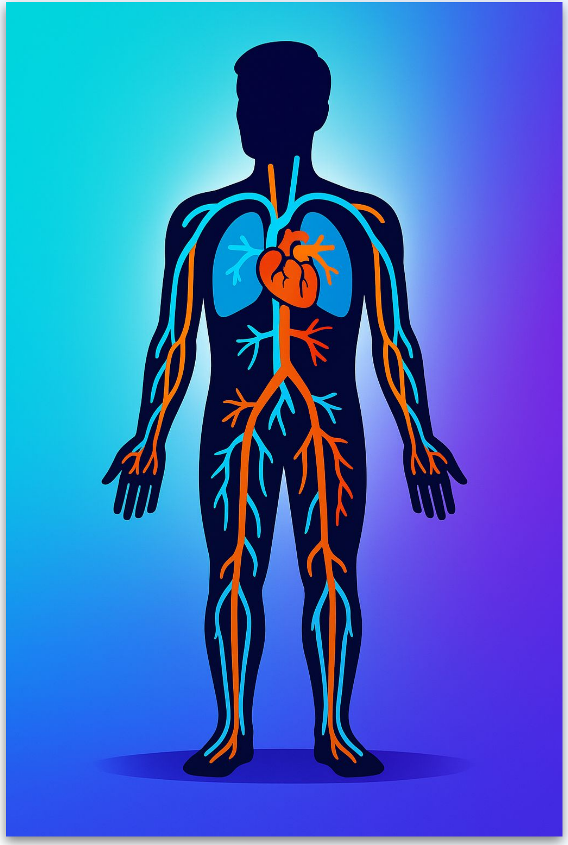
How good are today's specialized FMs?

Toto: An Observability TSFM

Datadog AI Research

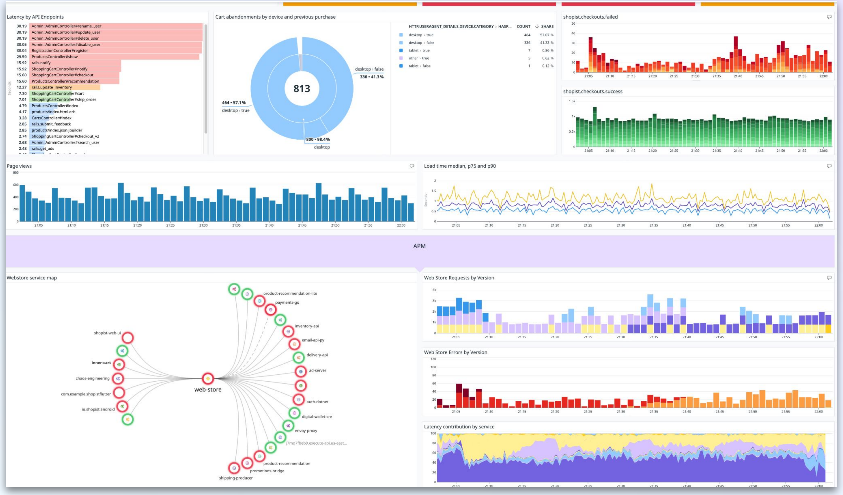
What is **Observability?**

Observing/monitoring: The Human Body



AI-generated Illustrations

Observing/monitoring: Computer Systems



Latency by API Endpoints

30.19	Admin::AdminController#rename_user
30.19	Admin::AdminController#update_user
30.19	Admin::AdminController#delete_user
30.05	Admin::AdminController#disable_user
30.04	RegistrationController#register
29.59	ProductsController#show
15.92	rails.notify
15.92	ShoppingCartController#notify
15.60	ShoppingCartController#checkout
15.60	ProductsController#recommendation
12.27	rails.update_inventory
7.30	ShoppingCartController#cart
7.01	ShoppingCartController#ship_order
4.79	ProductsController#index
4.17	products/index.html.erb
3.28	CartsController#index
2.85	rails.submit_feedback
2.85	products/index.json.jbuilder
2.74	ShoppingCartController#checkout_v2
2.68	Admin::AdminController#search_user
2.48	rails.get_ads

HTTP.USERAGENT_DETAILS.DEVICE.CATEGORY › HASP...	COUNT	↓ SHARE
desktop › true	464	57.07 %
desktop › false	336	41.33 %
tablet › true	7	0.86 %
other › true	5	0.62 %
tablet › false	1	0.12 %

AI-generated Illustrations

DATADOG

1,000s of hosts, pods, containers, etc.

Trillions

of data points/hour

Types of Observability Data

Telemetry Data

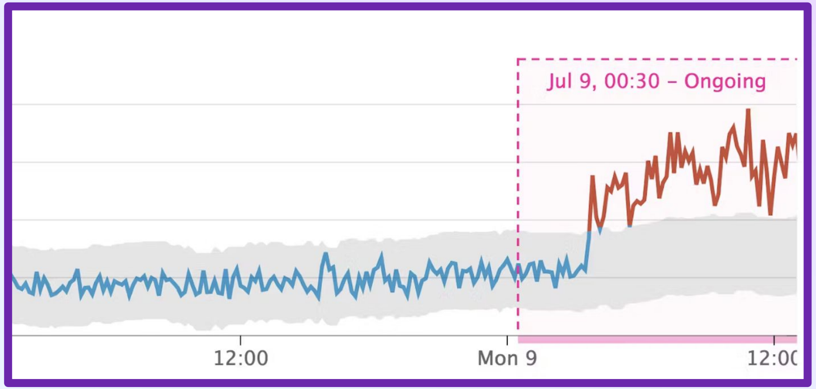
- Metrics
- Logs
- Traces
- Network Flows
- Source Code
- Cloud Cost
- CI / CD Pipelines
- Security Signals
- …

Human Interaction Data

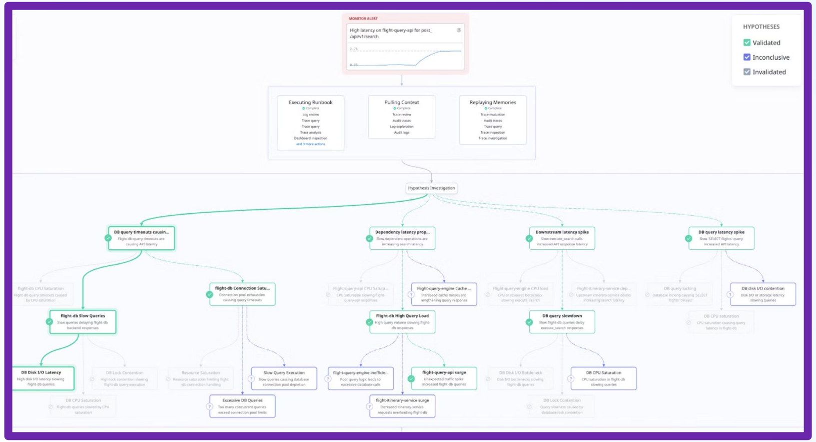
- Monitors configuration
- Dashboards configuration
- Notebooks configuration
- Interactive usage during an investigation
- …

AI Opportunities

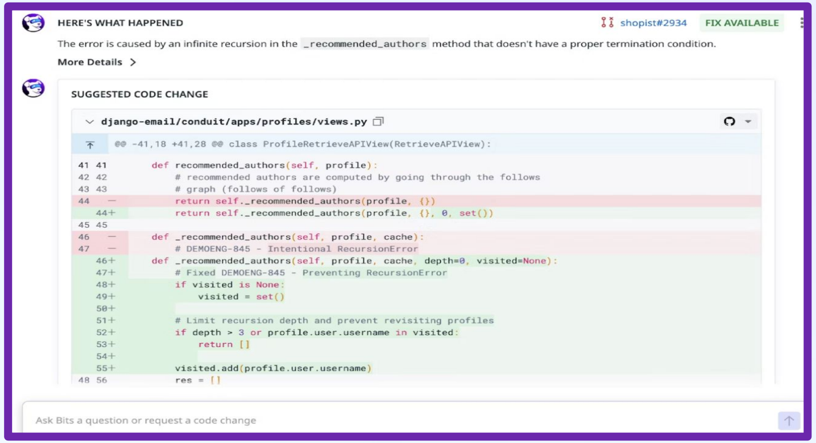
Forecasting & Anomaly Detection



Root Cause Analysis


MONITOR ALERT
High latency on flight-query-api for post_ /api/v1/search
HYPOTHESES
☑ Validated
☑ Inconclusive
☑ Invalidated
Executing Runbook
Complete
Log review
Trace query
Trace query
Trace analysis
Dashboard inspection
and 3 more actions
Pulling Context
Complete
Trace review
Audit traces
Log exploration
Audit logs
Replaying Memories
Complete
Trace evaluation
Audit traces
Trace query
Trace inspection
Trace investigation
Hypothesis Investigation
DB query timeouts causin...
Flight-db query timeouts are causing API latency
Dependency latency prop...
Slow dependent operations are increasing search latency
Downstream latency spike
Slow execute_search calls increased API response latency
DB query latency spike
Slow 'SELECT flights' query increased API latency
flight-db CPU Saturation
Flight-db query timeouts caused by CPU saturation
flight-db Connection Satu...
Connection pool exhaustion causing query timeouts
Flight-query-api CPU Satura...
CPU saturation slowing flight-query-api responses
Flight-query-engine Cache ...
Increased cache misses are lengthening query response
Flight-query-engine CPU load
CPU or resource bottleneck slowing execute_search
Flight-itinerary-service dep...
Upstream itinerary service delays increasing search latency
DB query locking
Database locking causing 'SELECT flights' delays?
DB disk I/O contention
Disk I/O or storage latency slowing queries
flight-db Slow Queries
Slow queries delaying flight-db backend responses
Flight-db High Query Load
High query volume slowing flight-db responses
DB query slowdown
Slow flight-db queries delay execute_search responses
DB CPU saturation
CPU saturation causing query latency in flight-db
DB Disk I/O Latency
High disk I/O latency slowing flight-db queries
DB Lock Contention
High lock contention slowing flight-db query execution
Resource Saturation
Resource saturation limiting flight-db connection handling
Slow Query Execution
Slow queries causing database connection pool depletion
flight-query-engine inefficie...
Poor query logic leads to excessive database calls
flight-query-api surge
Unexpected traffic spike increased flight-db queries
DB Disk I/O Bottleneck
Disk I/O bottlenecks slowing flight-db queries
DB CPU Saturation
CPU saturation in flight-db slowing queries
DB CPU Saturation
Flight-db queries slowed by CPU saturation
Excessive DB Queries
Too many concurrent queries exceed connection pool limits
flight-itinerary-service surge
Increased itinerary-service requests overloading flight-db
DB Lock Contention
Query slowness caused by database lock contention

Production Code Repair



HERE'S WHAT HAPPENED

The error is caused by an infinite recursion in the `_recommended_authors` method that doesn't have a proper termination condition.

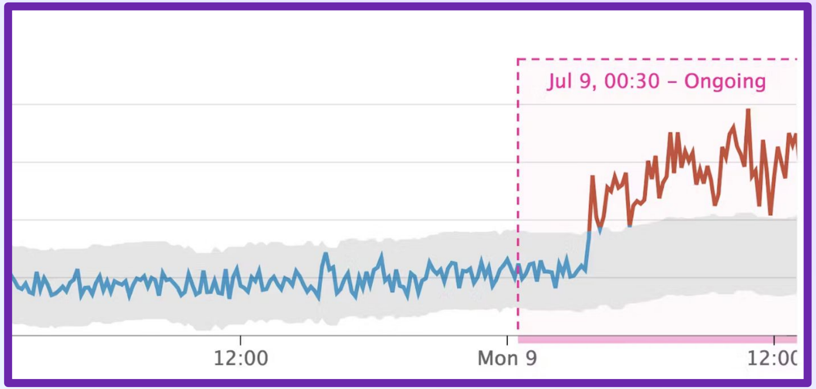
SUGGESTED CODE CHANGE

django-email/conduit/apps/profiles/views.py

```
@@ -41,18 +41,28 @@ class ProfileRetrieveAPIView(RetrieveAPIView):
41 41      def recommended_authors(self, profile):
42 42          # recommended authors are computed by going through the follows
43 43          # graph (follows of follows)
44    —         return self._recommended_authors(profile, {})
   44+         return self._recommended_authors(profile, {}, 0, set())
45 45
46    —     def _recommended_authors(self, profile, cache):
47    —         # DEMOENG-845 - Intentional RecursionError
   46+     def _recommended_authors(self, profile, cache, depth=0, visited=None):
   47+         # Fixed DEMOENG-845 - Preventing RecursionError
   48+         if visited is None:
   49+             visited = set()
   50+
   51+         # Limit recursion depth and prevent revisiting profiles
   52+         if depth > 3 or profile.user.username in visited:
   53+             return []
   54+
   55+         visited.add(profile.user.username)
48 56          res = []
```

Can we just apply existing Time Series FMs (TSFMs)?

Forecasting & Anomaly Detection

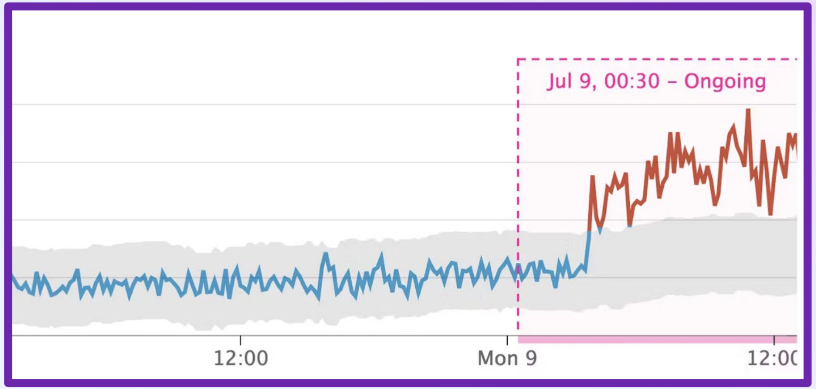


Promise:

- Several models in recent years
- Zero shot capabilities

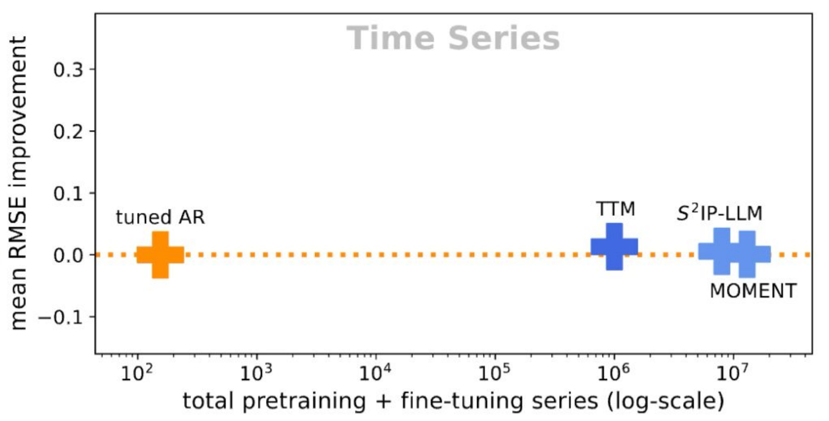
Can we just apply existing Time Series FMs (TSFMs)?

Forecasting & Anomaly Detection



Issues:

- Don't beat supervised baselines
- Not tailored to observability

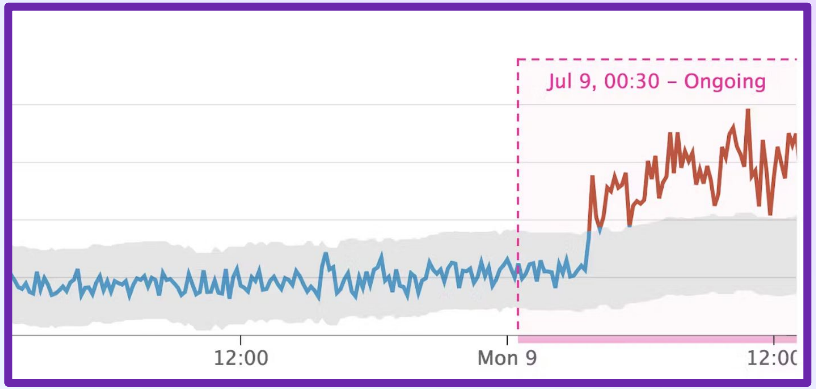


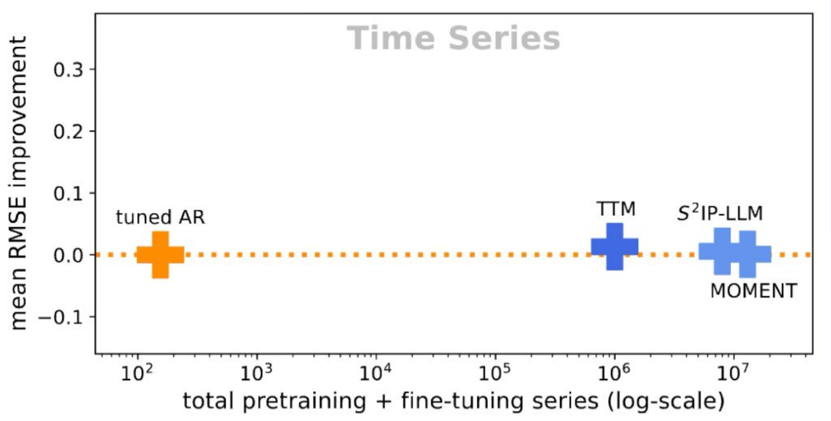
DATADOG

Can we just apply existing Time Series FMs (TSFMs)?

Our work: specialize eval, data & modeling for Observability!

Forecasting & Anomaly Detection





DATADOG

BOOM

New Observability Benchmark

Largest time series benchmark

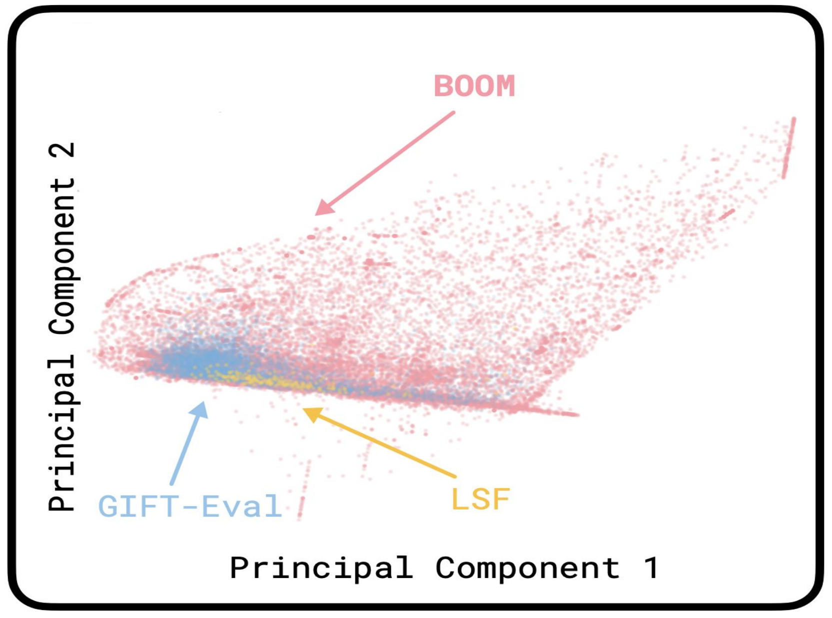
Comprised of Real Data

Internal observability data from Datadog

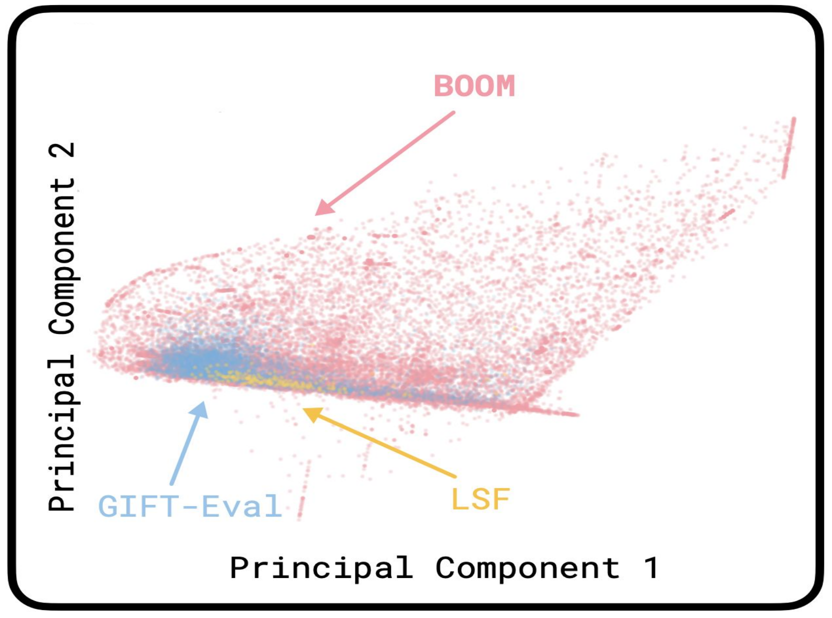
Open Source

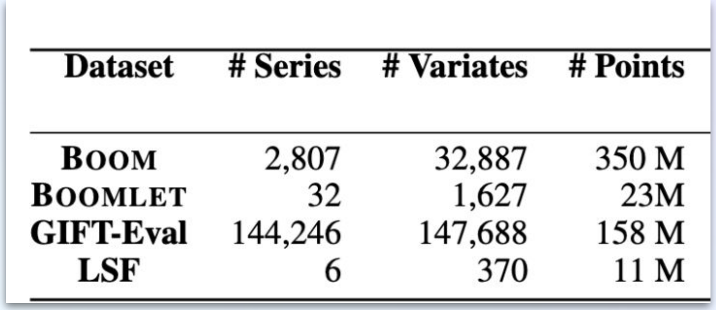
Apache 2.0

27K HF downloads



Captures challenge of real-world observability data





Dataset	# Series	# Variates	# Points
BOOM	2,807	32,887	350 M
BOOMLET	32	1,627	23M
GIFT-Eval	144,246	147,688	158 M
LSF	6	370	11 M

Toto

Time Series Foundation Model

150M param decoder-only architecture

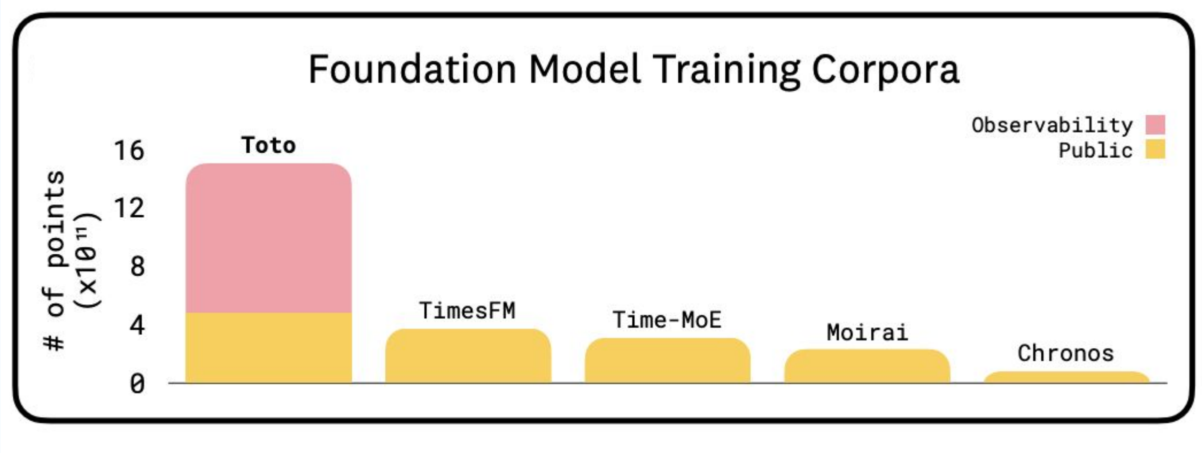
Optimized for Observability

And also SOTA on general-purpose time series forecasting

Open Weights

Apache 2.0

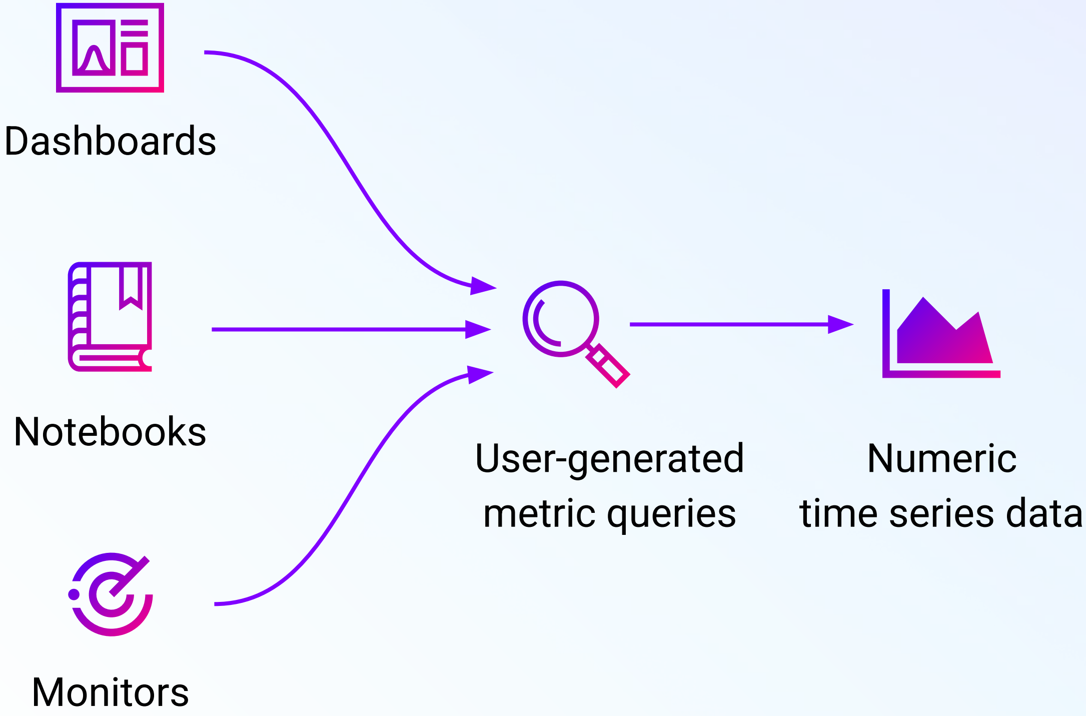
~8M HF downloads



(Datadog internal data only)

DATADOG

Data Collection (Datadog internal data only)



Dashboards

Notebooks

Monitors

User-generated metric queries

Numeric time series data

Collect each query over:

- Multiple time slices
- Different time intervals

High-cardinality multivariate data

Proportional Attention: judiciously attend across covariates

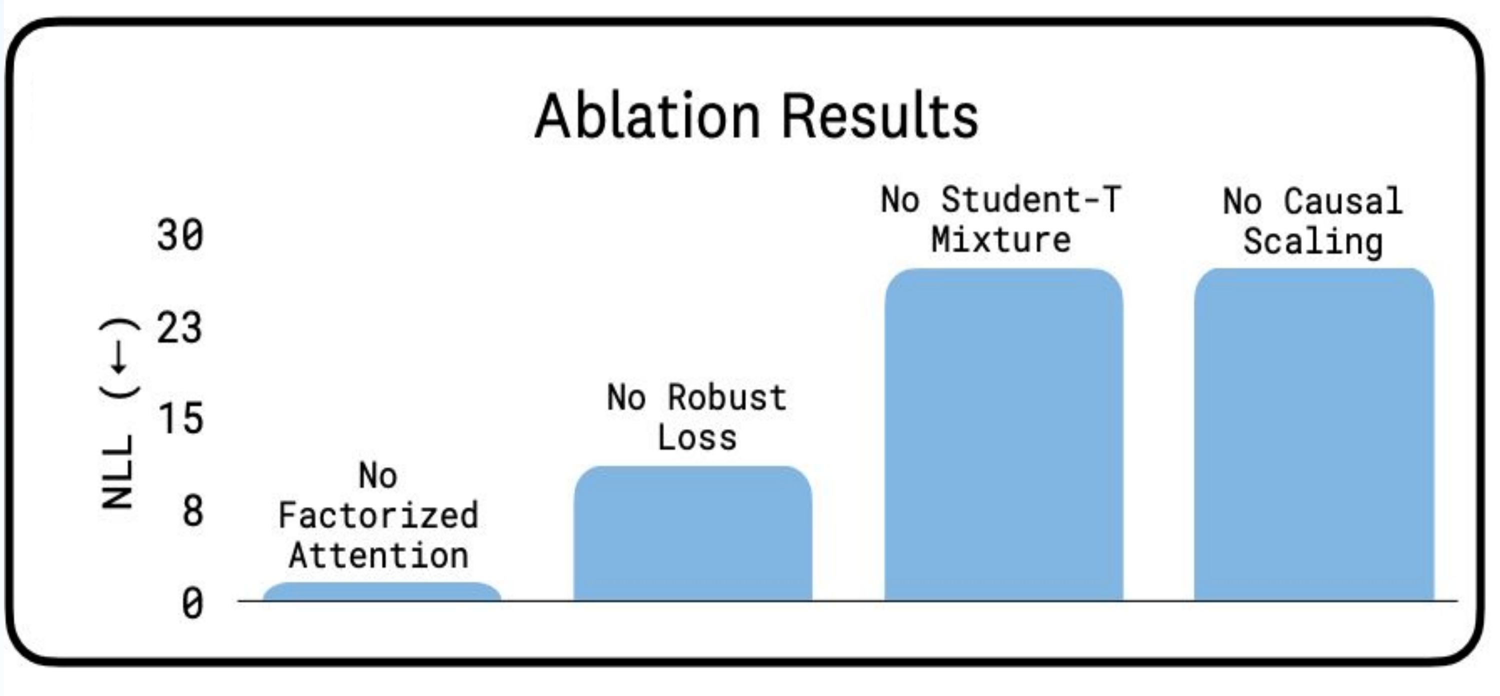
Skewed, heavy tailed distributions

Student-T mixture & robust loss: for improved modeling and learning

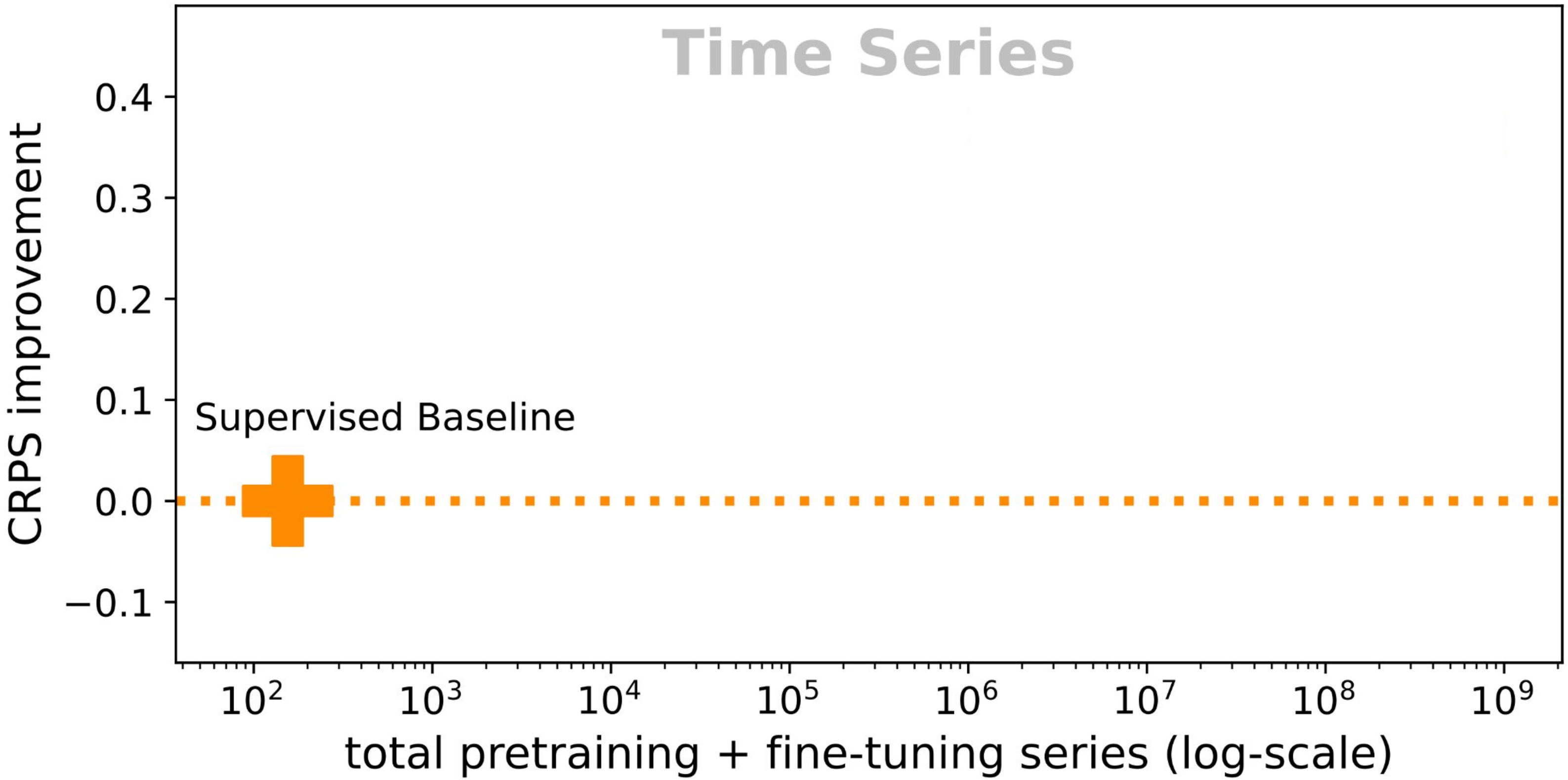
Extreme dynamic range, nonstationarity

Patch-based causal scaling: address highly non-stationary TS

These modifications make a big difference



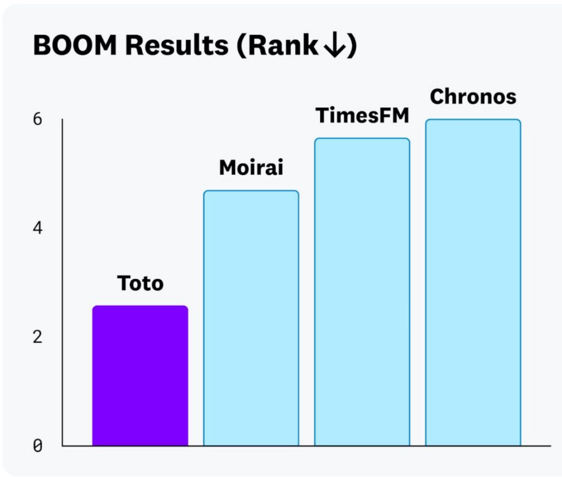
BOOM Results

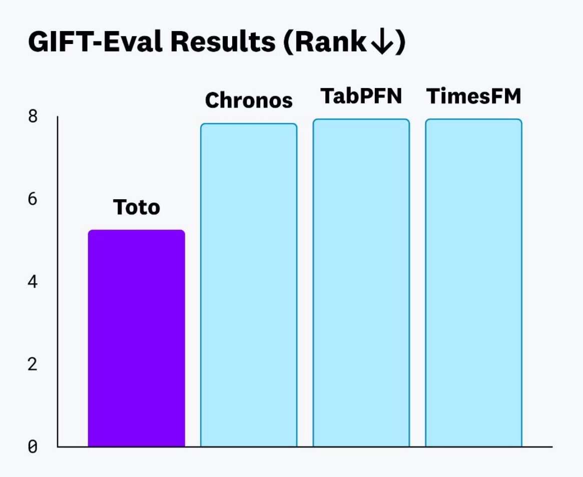


Specialized Observability FMs

Datadog AI Research

[Cohen*-Khwaja*-et al.]





TSFMs have achieved their "BERT moment"!

Specialization matters

At time of release

Ongoing Work

Product Applications

Multimodality

Scaling

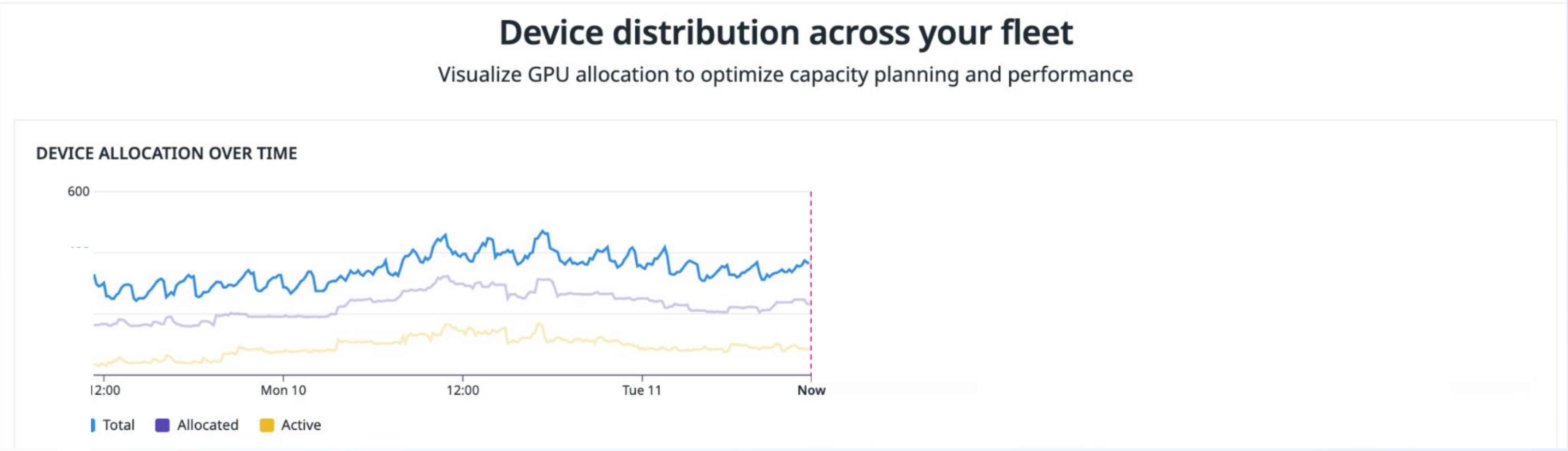
Ongoing Work

Product Applications

Multimodality

Scaling

GPU Monitoring: how many GPUs will I need?



Device distribution across your fleet

Visualize GPU allocation to optimize capacity planning and performance

Important for budgeting/planning

Mature production tool already exists

Natural application of Zero Shot Toto

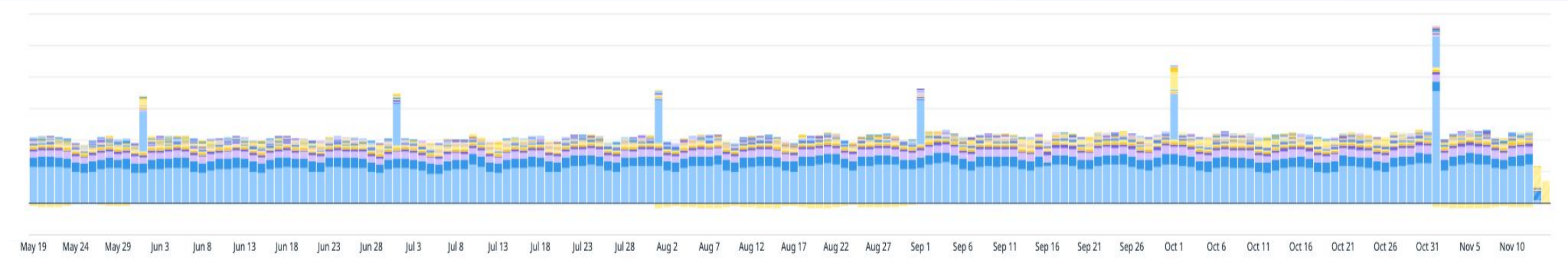
How does Toto perform?

No clear winner between ZS Toto and mature production tool

ZS Toto preferred by humans in 71% of cases in blind evaluation

In production as of last month!

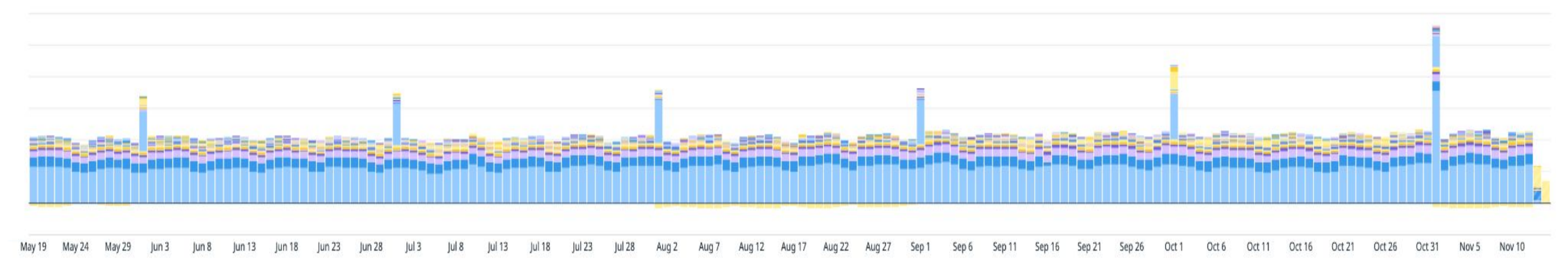
Forecasting Cloud Costs: how much $ will I spend?



Important for budgeting/planning

Also a seemingly natural application of Zero Shot Toto

Forecasting Cloud Costs: how much $ will I spend?

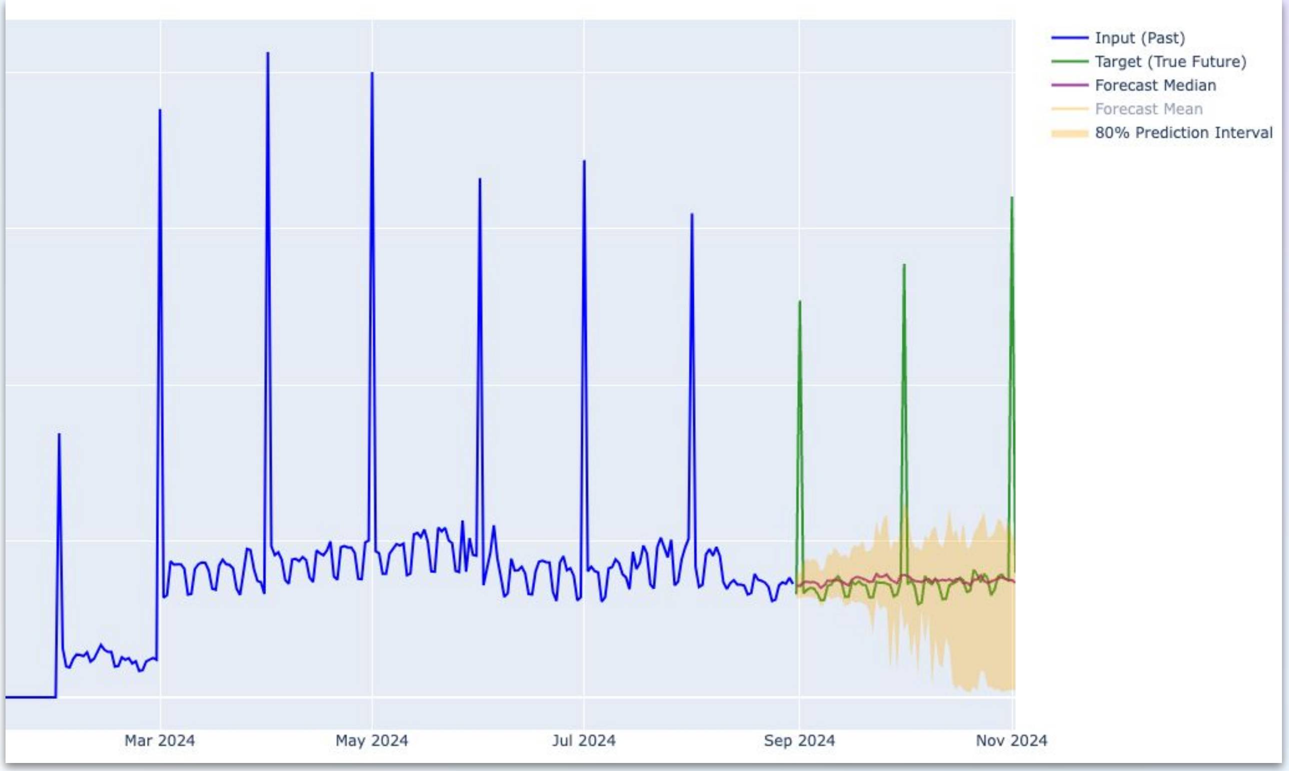


Important for budgeting/planning

Challenges:

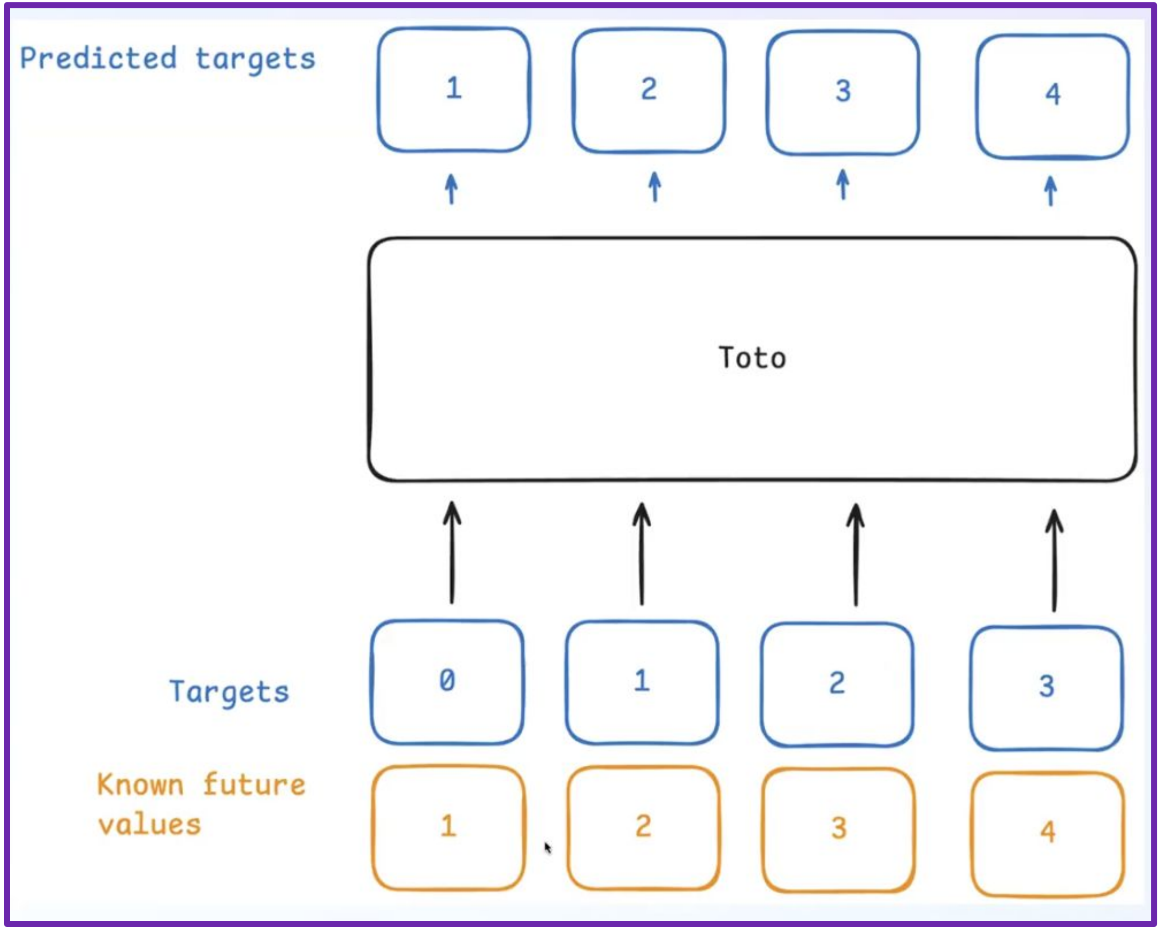
- Large, irregular seasonality effects (e.g. day-of-month)
- Not much historical data available
- Product-specific eval differs from typical TSFM benchmarking
- Strong latency constraints

Zero shot Toto misses the spikes



Especially tricky b/c of uneven period lengths

Provide Toto with ‘exogenous variable’, e.g., day of month!


Predicted targets
1
2
3
4
Toto
Targets
0
1
2
3
Known future values
1
2
3
4

Exogenous variable fine-tuning

Input preprocessing

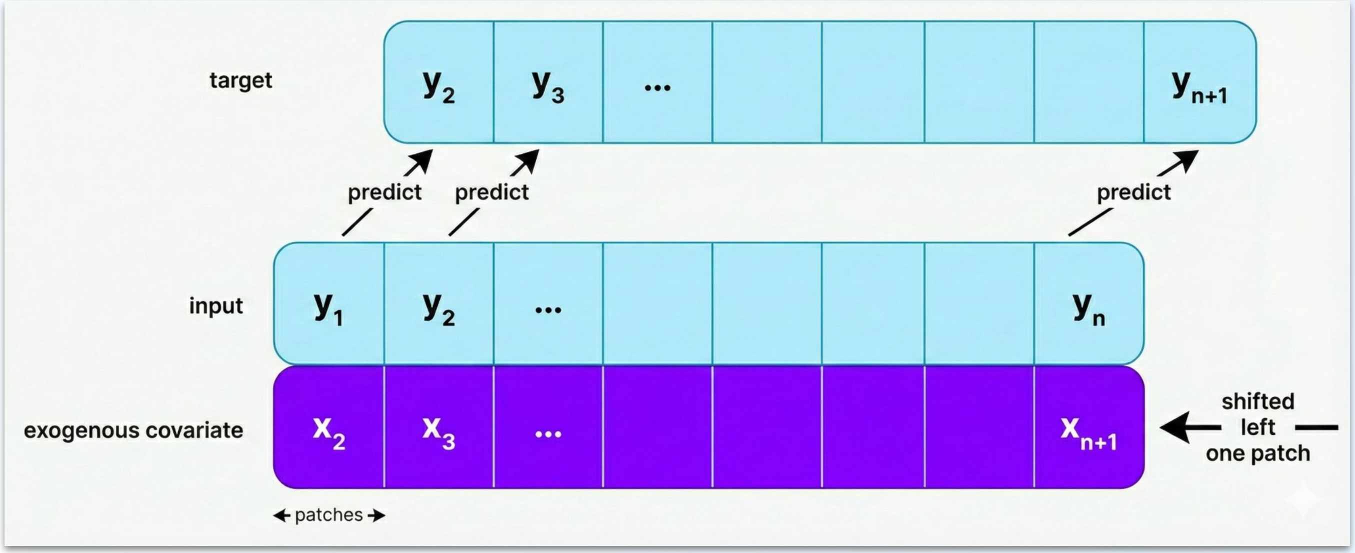
- Use dummy variable for first of the month
- Shifted exogenous variables one patch into the future
- Stacked along variate dimension

Training

- Mask loss for exogenous variable

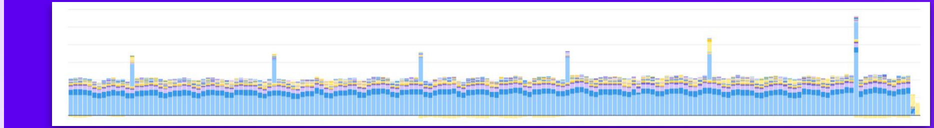
Inference

- Inject known future exogenous values during decoding



DATADOG

Forecasting Cloud Costs



CHALLENGES

Irregular seasonality effects

Product-imposed latency constraints

Product-specific evaluation

SOLUTIONS

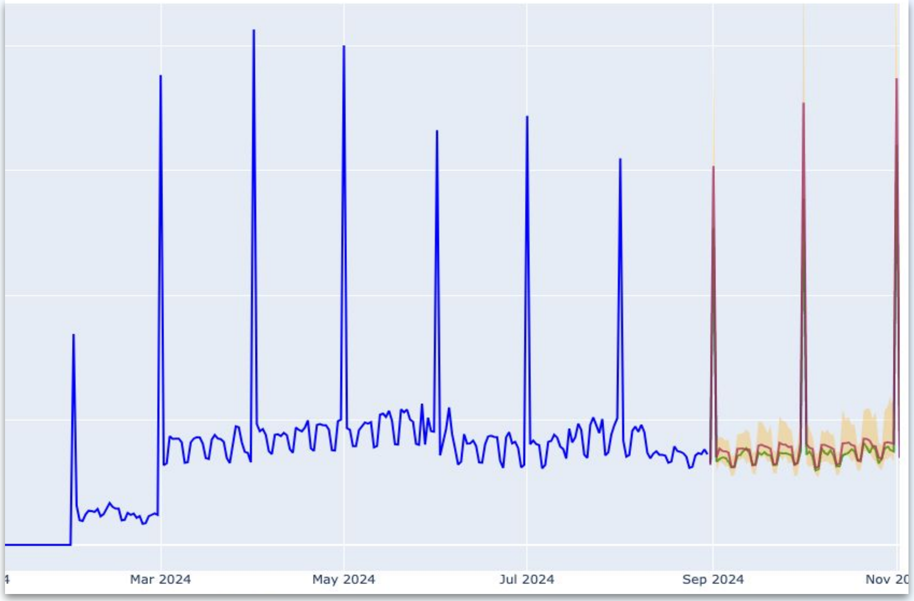
Toto + FT + Exogenous variables

Toto inference meets latency requirements

New benchmark & metrics

DATADOG

The end result...



17% improvement over baseline

41% over zero-shot

Satisfies latency reqs

Coming soon: Support for fine-tuning and exogenous variables
https://github.com/DataDog/toto

Toto applications, next steps

Autoscaling – Forecast demand so services can right-size

Predictive alerting – Forecast issues before they happen (and ideally fix them w/o needing to page an engineer)

Datadog AI Research:
Vertical AI for observability

We are hiring!