ICLR

International Conference On
Learning Representations

ICLR

International Conference On
Learning Representations

Observability Foundation Models

From Scaling Time Series Foundation Models to Multimodal “World Models”

ICLR 2026
Othmane Abou-Amal – Director, Datadog AI Research

DATADOG

Observability Foundation Models

From Scaling Time Series Foundation Models to Multimodal "World Models"

ICLR 2026

Othmane Abou-Amal – Director, Datadog AI Research

DATADOG

Observability Foundation Models

From Scaling Time Series Foundation Models to Multimodal “World Models”

ICLR 2026
Othmane Abou-Amal – Director, Datadog AI Research

DATADOG

Observability Foundation Models

From Scaling Time Series Foundation Models to Multimodal “World Models”

ICLR 2026
Othmane Abou-Amal – Director, Datadog AI Research

DATADOG

Observability Foundation Models

From Scaling Time Series Foundation Models to Multimodal “World Models”

ICLR 2026

Othmane Abou-Amal – Director, Datadog AI Research

Observability Foundation Models

From Scaling Time Series Foundation Models to Multimodal “World Models”

ICLR 2026

Othmane Abou-Amal – Director, Datadog AI Research

DATADOG

Agenda

01 Observability 101

02 Why Foundation Models for observability

03 Scaling Toto - From the Bert moment to the GPT-2 Moment

04 From TSFMs to Multimodal “World Models”

05 Questions?

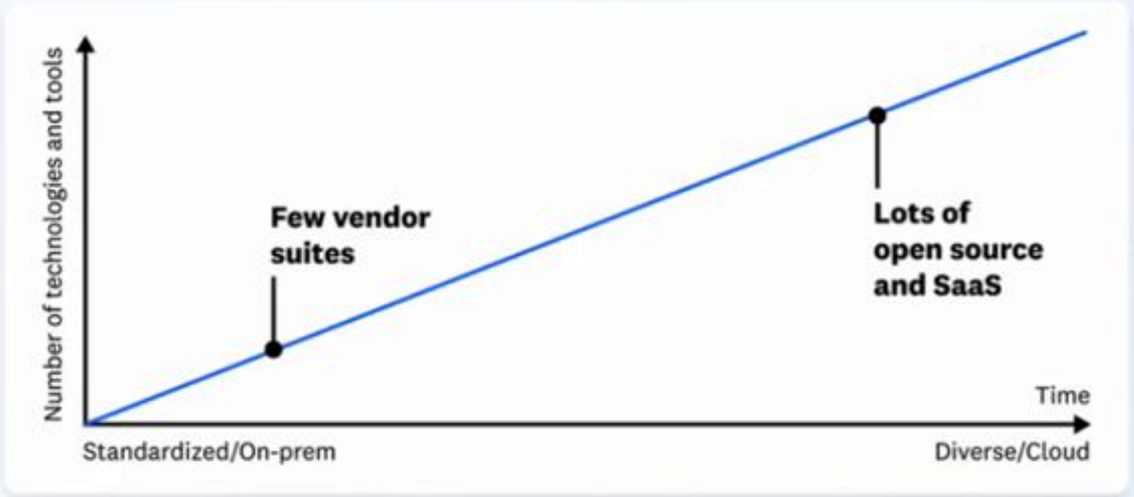
Observability 101

The **3am** page problem!

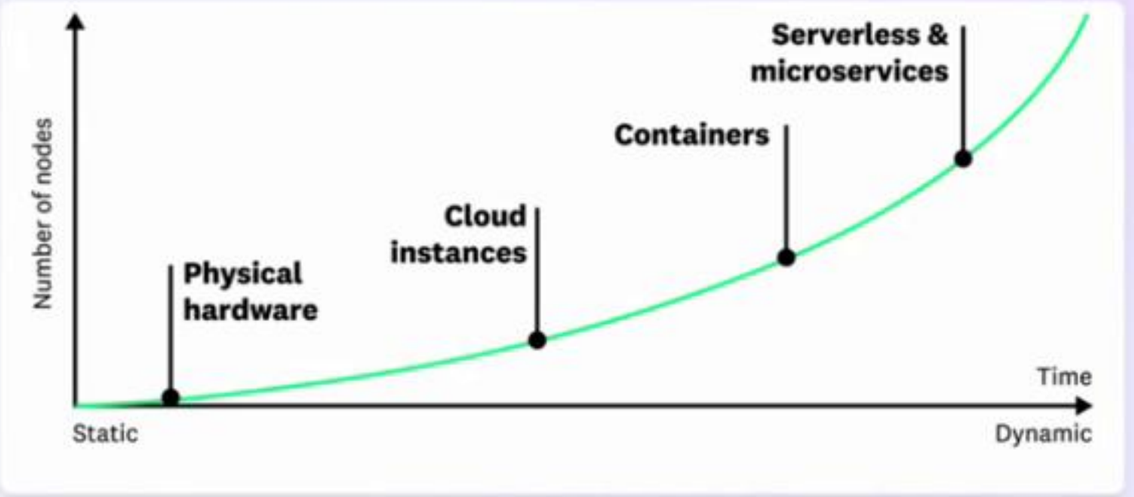
The **3am** page problem!

The problem: an explosion of complexity

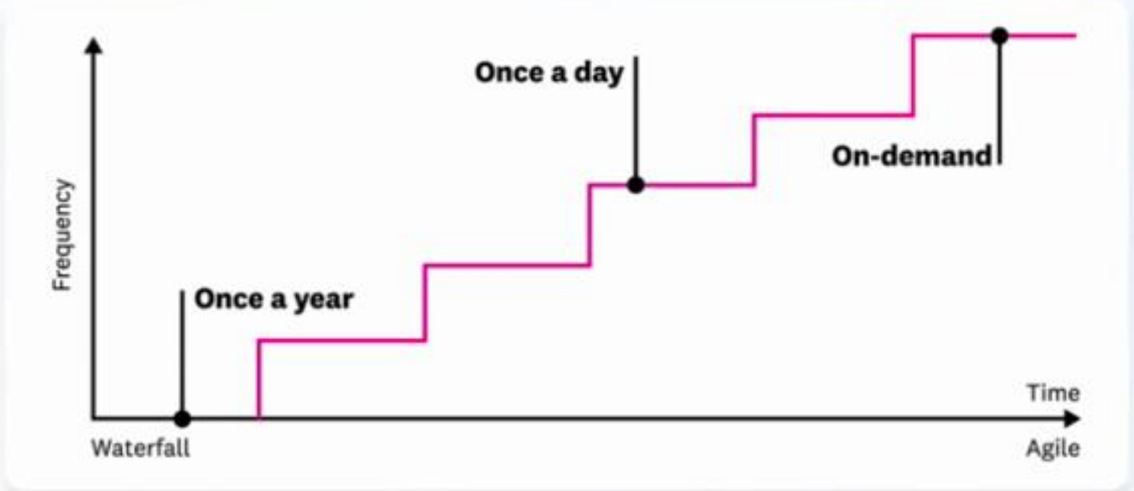
Diversity of technologies in use



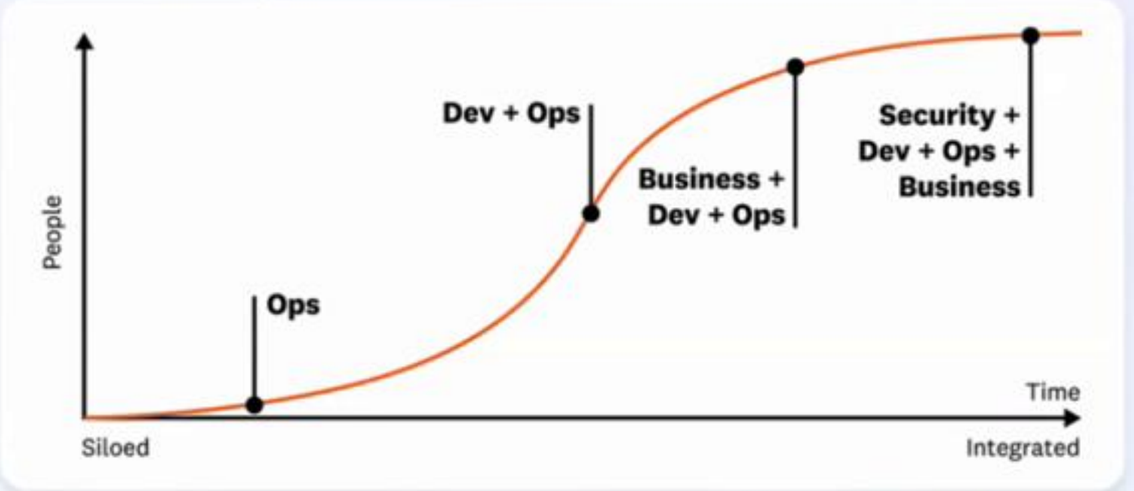
Scale in number of computing units



Frequency of release

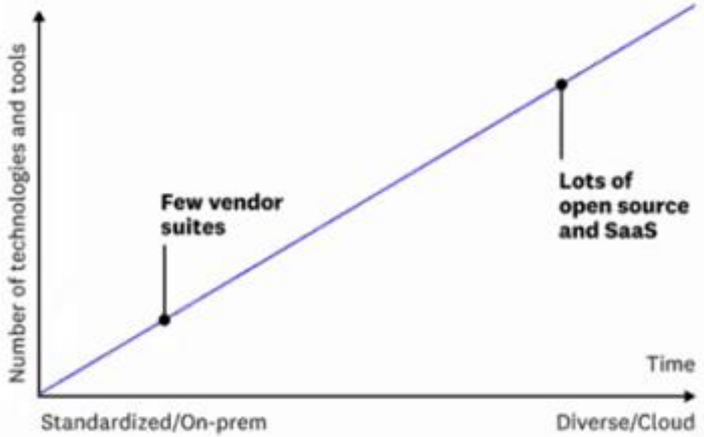


Number of people involved

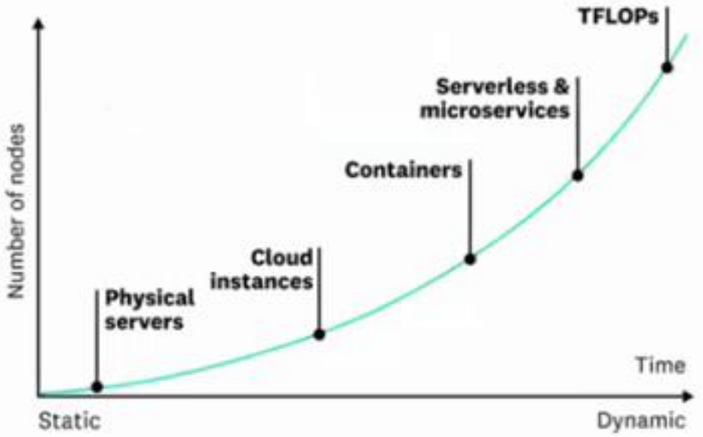


AI compounds complexity

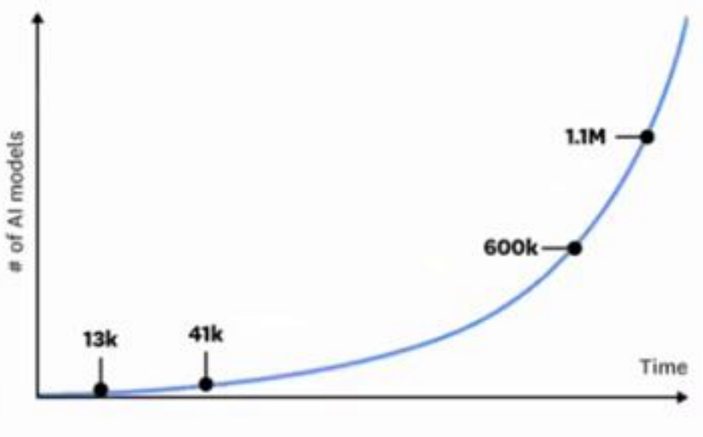
Diversity of technologies in use



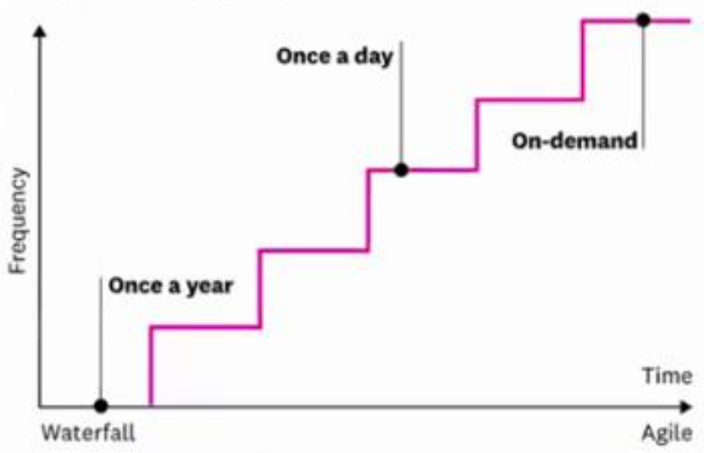
Scale in number of computing units



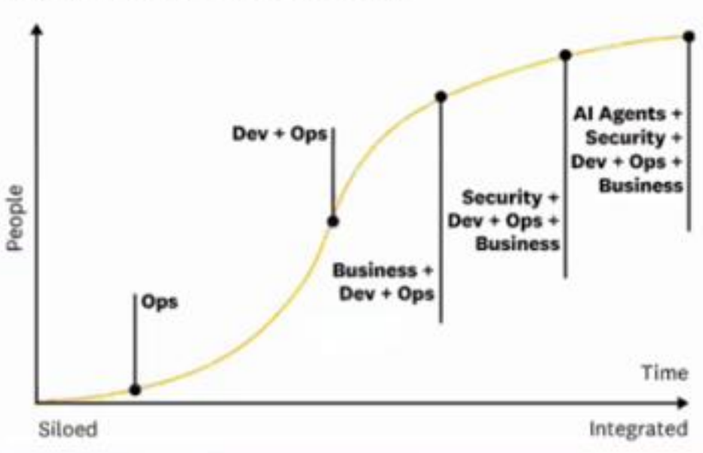
Number of Models on Hugging Face



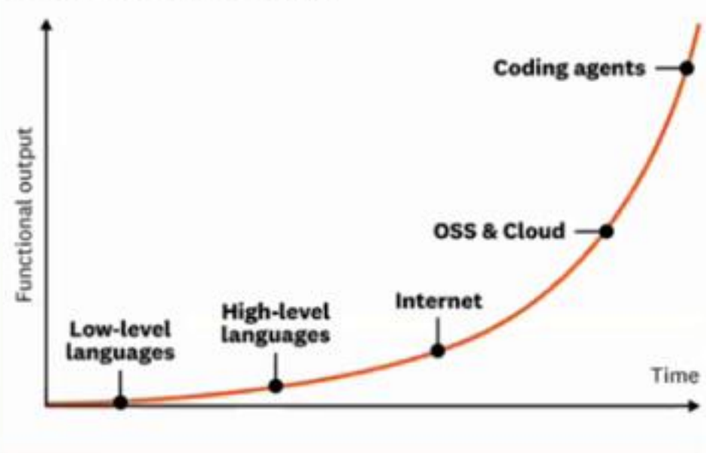
Frequency of releases



Number of people involved



Developer productivity

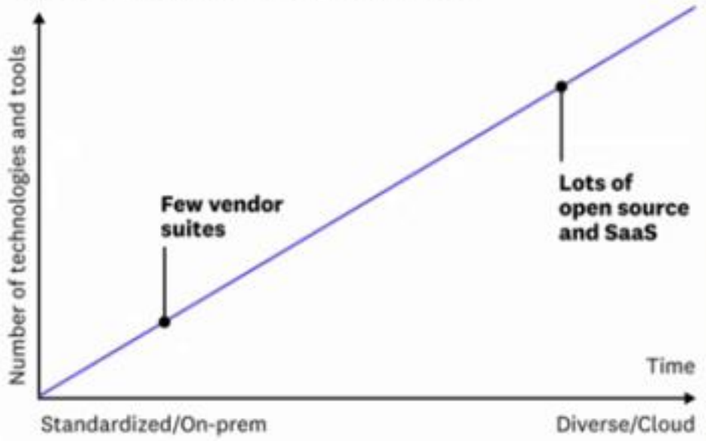


Source for number of models: Hugging Face Hub Stats Dashboard, cfahlgren1, 2025.

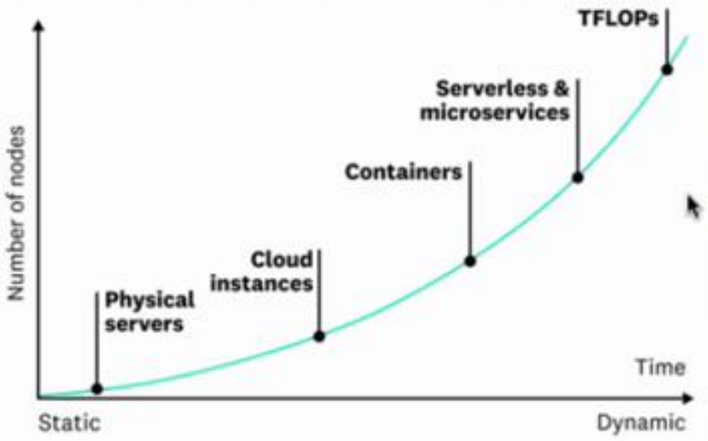
DATADOG

AI compounds complexity

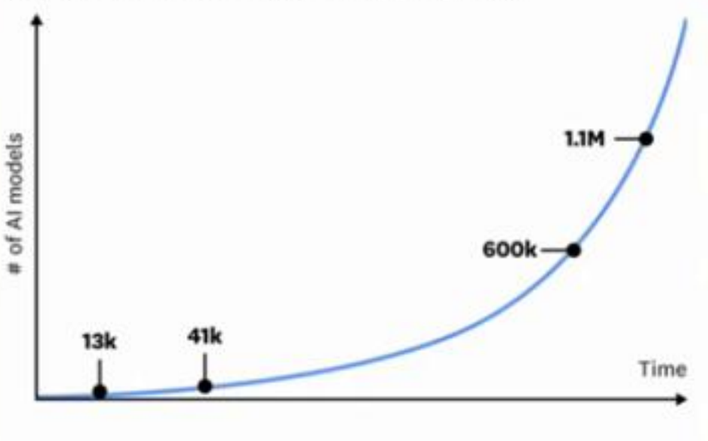
Diversity of technologies in use



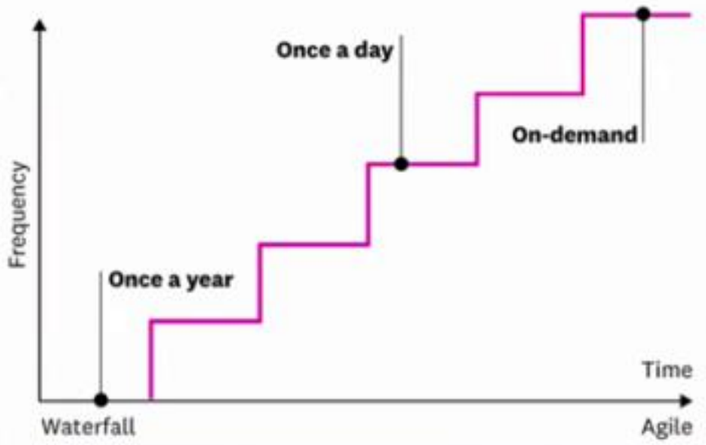
Scale in number of computing units



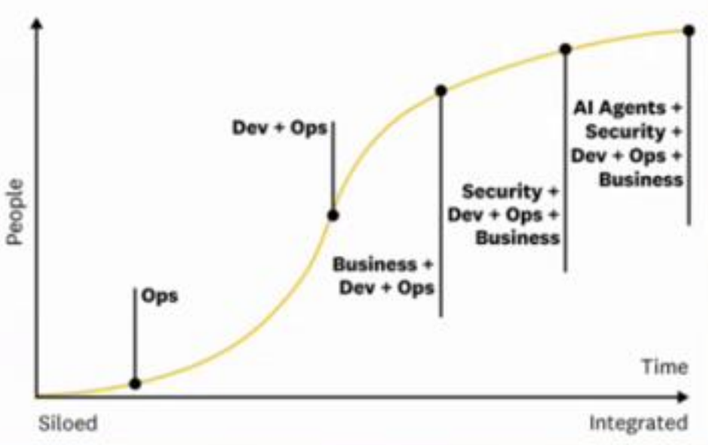
Number of Models on Hugging Face



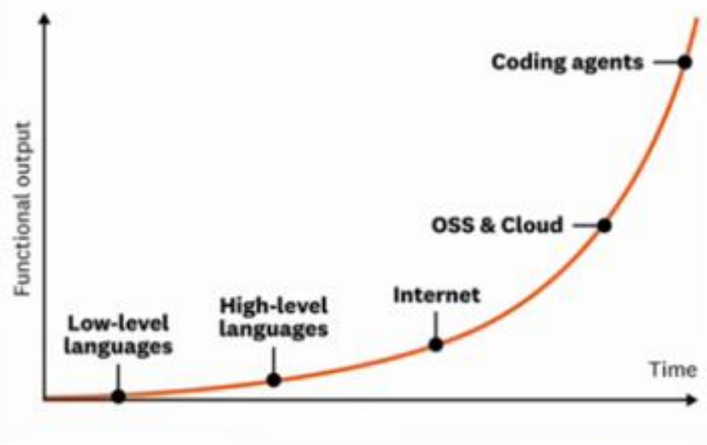
Frequency of releases



Number of people involved



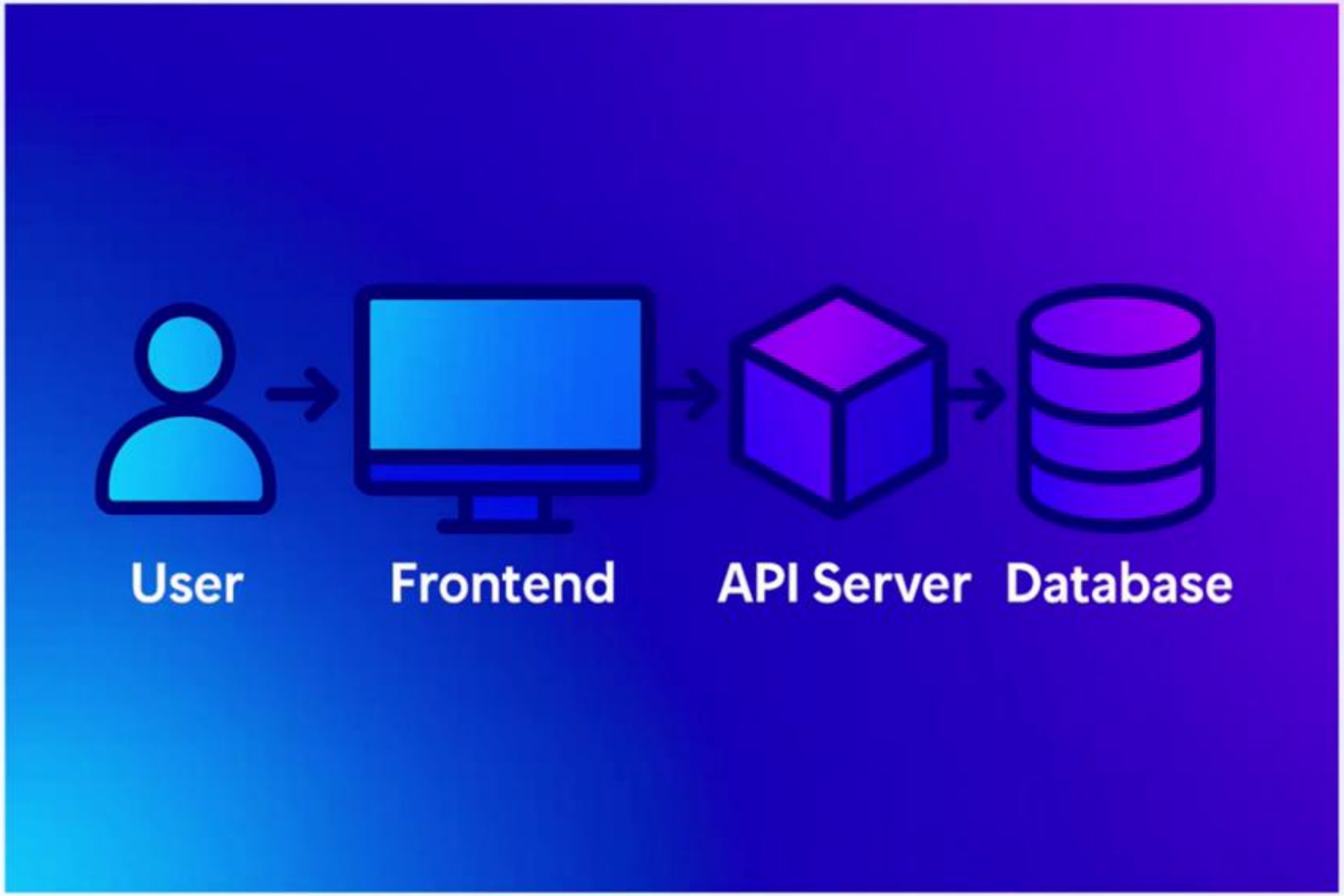
Developer productivity



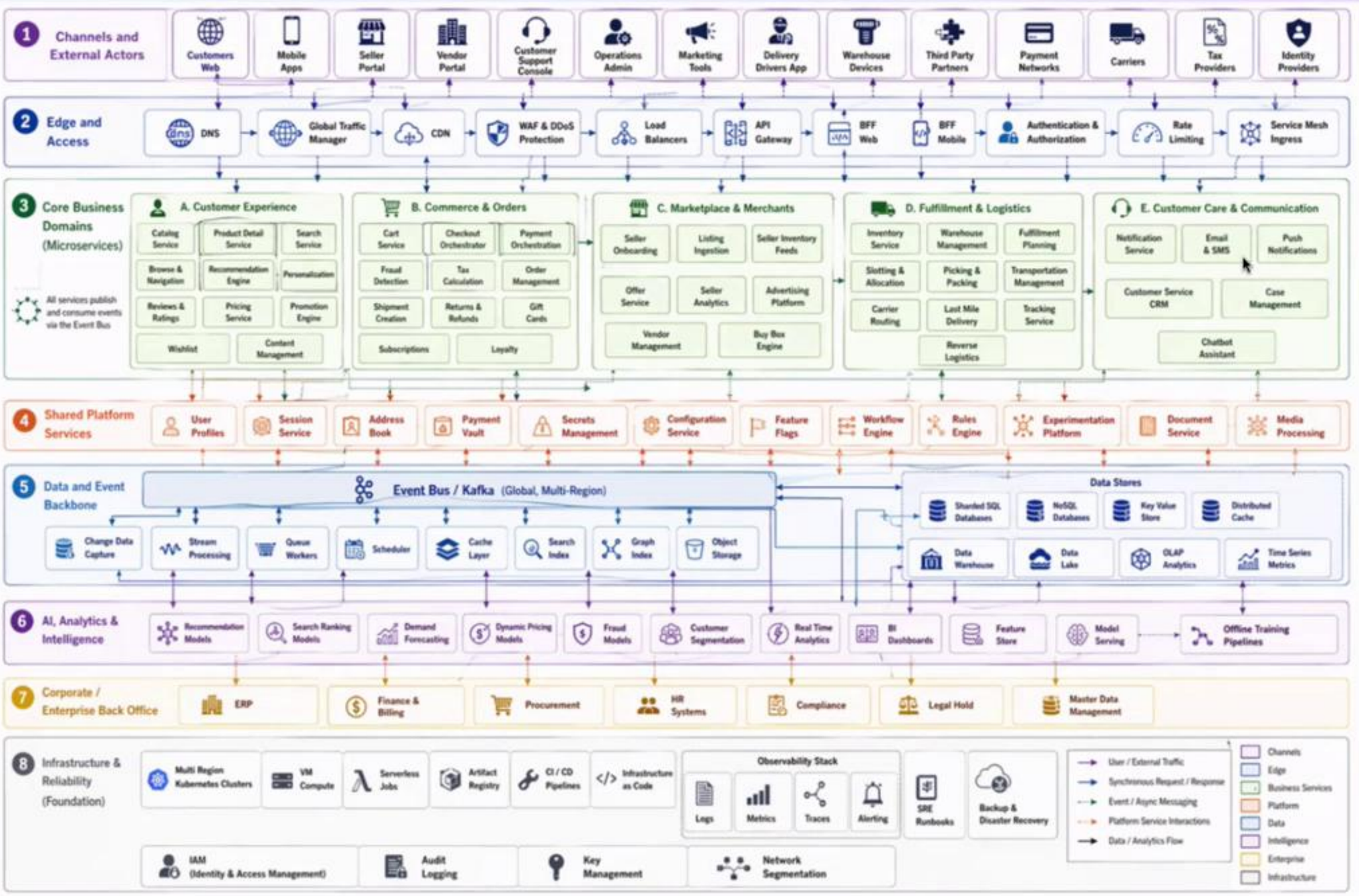
Source for number of models: Hugging Face Hub Stats Dashboard, cfahlgren1, 2025.

DATADOG

Example: A “simple” e-commerce site



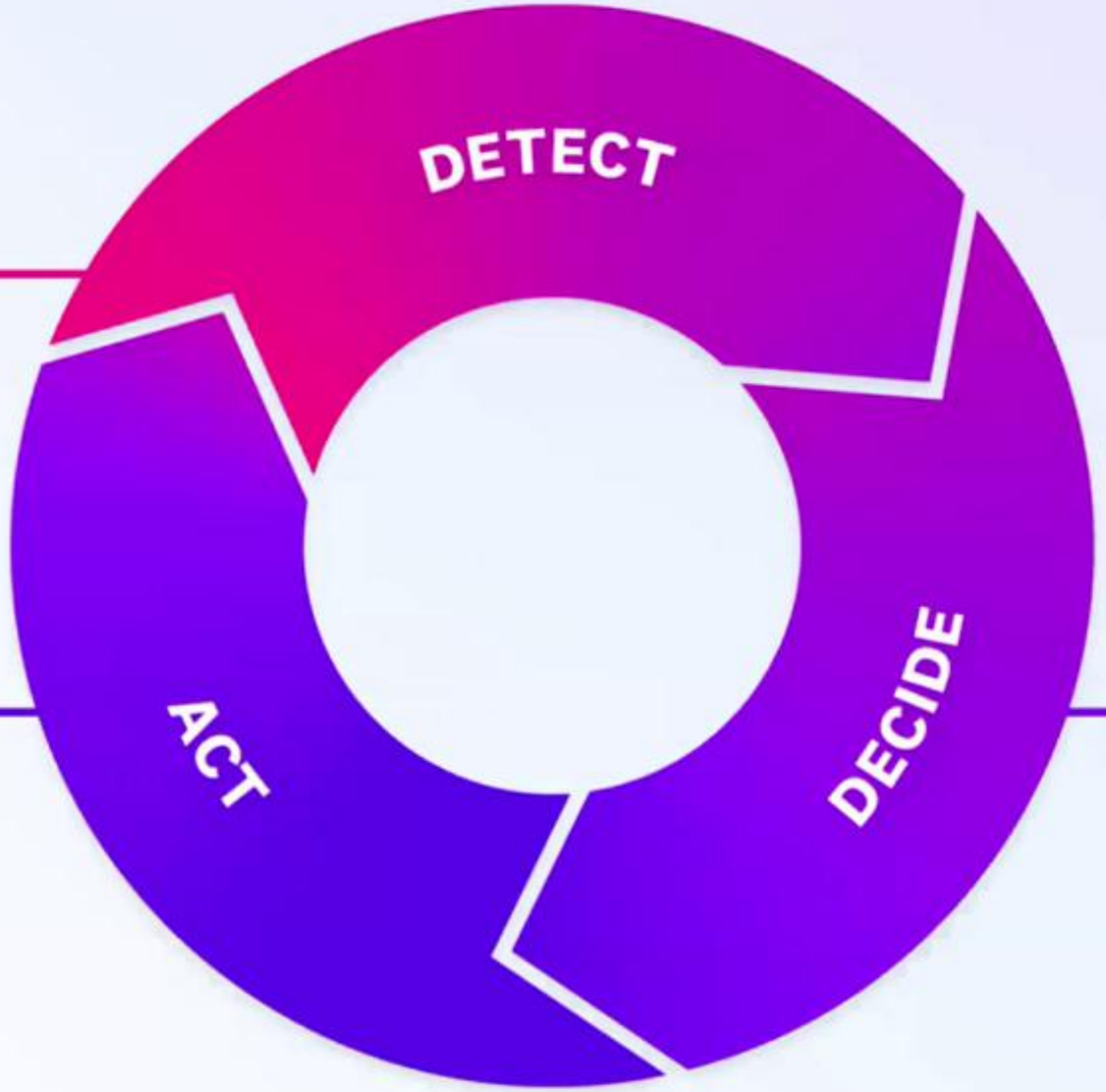
Exa


1 Channels and External Actors
Customers Web
Mobile Apps
Seller Portal
Vendor Portal
Customer Support Console
Operations Admin
Marketing Tools
Delivery Drivers App
Warehouse Devices
Third Party Partners
Payment Networks
Carriers
Tax Providers
Identity Providers
2 Edge and Access
DNS
Global Traffic Manager
CDN
WAF & DDoS Protection
Load Balancers
API Gateway
BFF Web
BFF Mobile
Authentication & Authorization
Rate Limiting
Service Mesh Ingress
3 Core Business Domains (Microservices)
All services publish and consume events via the Event Bus
A. Customer Experience
Catalog Service
Product Detail Service
Search Service
Browse & Navigation
Recommendation Engine
Personalization
Reviews & Ratings
Pricing Service
Promotion Engine
Wishlist
Content Management
B. Commerce & Orders
Cart Service
Checkout Orchestrator
Payment Orchestration
Fraud Detection
Tax Calculation
Order Management
Shipment Creation
Returns & Refunds
Gift Cards
Subscriptions
Loyalty
C. Marketplace & Merchants
Seller Onboarding
Listing Ingestion
Seller Inventory Feeds
Offer Service
Seller Analytics
Advertising Platform
Vendor Management
Buy Box Engine
D. Fulfillment & Logistics
Inventory Service
Warehouse Management
Fulfillment Planning
Slotting & Allocation
Picking & Packing
Transportation Management
Carrier Routing
Last Mile Delivery
Tracking Service
Reverse Logistics
E. Customer Care & Communication
Notification Service
Email & SMS
Push Notifications
Customer Service CRM
Case Management
Chatbot Assistant
4 Shared Platform Services
User Profiles
Session Service
Address Book
Payment Vault
Secrets Management
Configuration Service
Feature Flags
Workflow Engine
Rules Engine
Experimentation Platform
Document Service
Media Processing
5 Data and Event Backbone
Event Bus / Kafka (Global, Multi-Region)
Change Data Capture
Stream Processing
Queue Workers
Scheduler
Cache Layer
Search Index
Graph Index
Object Storage
Data Stores
Sharded SQL Databases
NoSQL Databases
Key Value Store
Distributed Cache
Data Warehouse
Data Lake
OLAP Analytics
Time Series Metrics
6 AI, Analytics & Intelligence
Recommendation Models
Search Ranking Models
Demand Forecasting
Dynamic Pricing Models
Fraud Models
Customer Segmentation
Real Time Analytics
BI Dashboards
Feature Store
Model Serving
Offline Training Pipelines
7 Corporate / Enterprise Back Office
ERP
Finance & Billing
Procurement
HR Systems
Compliance
Legal Hold
Master Data Management
8 Infrastructure & Reliability (Foundation)
Multi Region Kubernetes Clusters
VM Compute
Serverless Jobs
Artifact Registry
CI / CD Pipelines
Infrastructure as Code
Observability Stack
Logs
Metrics
Traces
Alerting
SRE Runbooks
Backup & Disaster Recovery
IAM (Identity & Access Management)
Audit Logging
Key Management
Network Segmentation
User / External Traffic
Synchronous Request / Response
Event / Async Messaging
Platform Service Interactions
Data / Analytics Flow
Channels
Edge
Business Services
Platform
Data
Intelligence
Enterprise
Infrastructure

DATADOG

Close the DevOps loop

Monitor systems for stress



When stress is detected, decide what actions to take

Take actions that will bring system back to health

Why Foundation Models for observability

This not a solved problem.

But

Data is our differentiated advantage

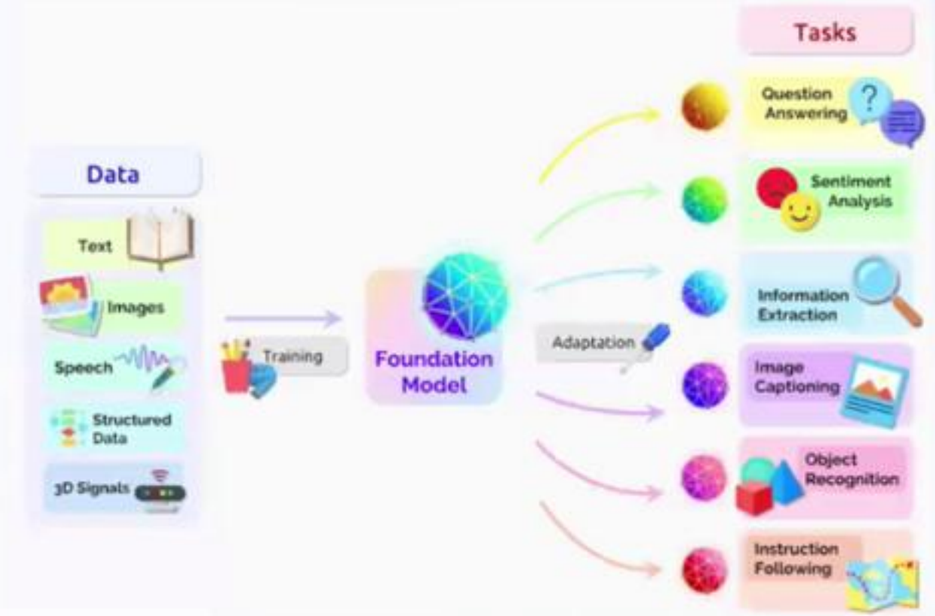
Trillions of datapoints per hour

Trillions of metrics

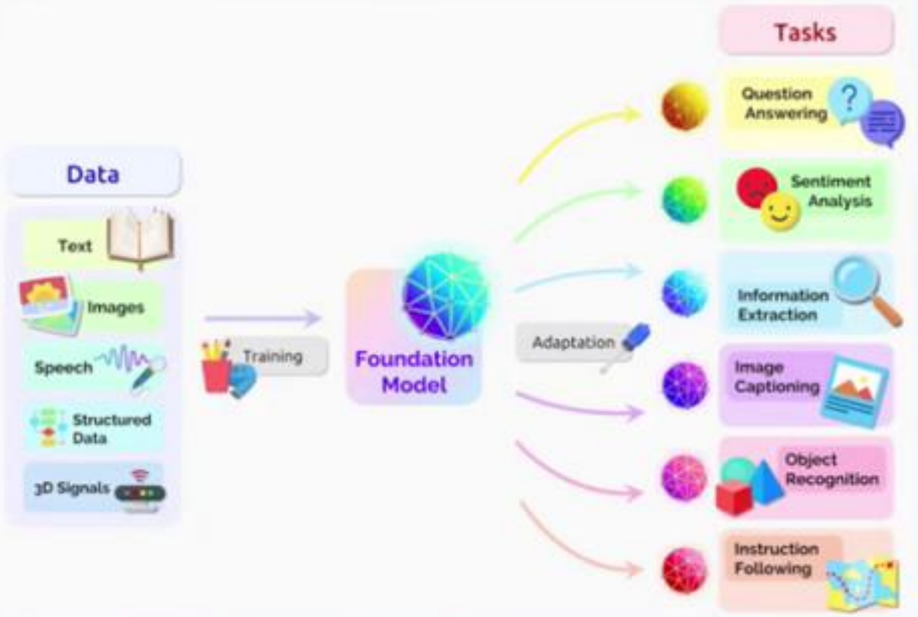
Billions of application traces

Exabytes of logs

Foundation Models unlock scale and performance across modalities



Foundation Models unlock scale and performance across modalities



Observability is next?

Toto

Time Series Foundation Model (TSFM)

SOTA zero-shot forecasts

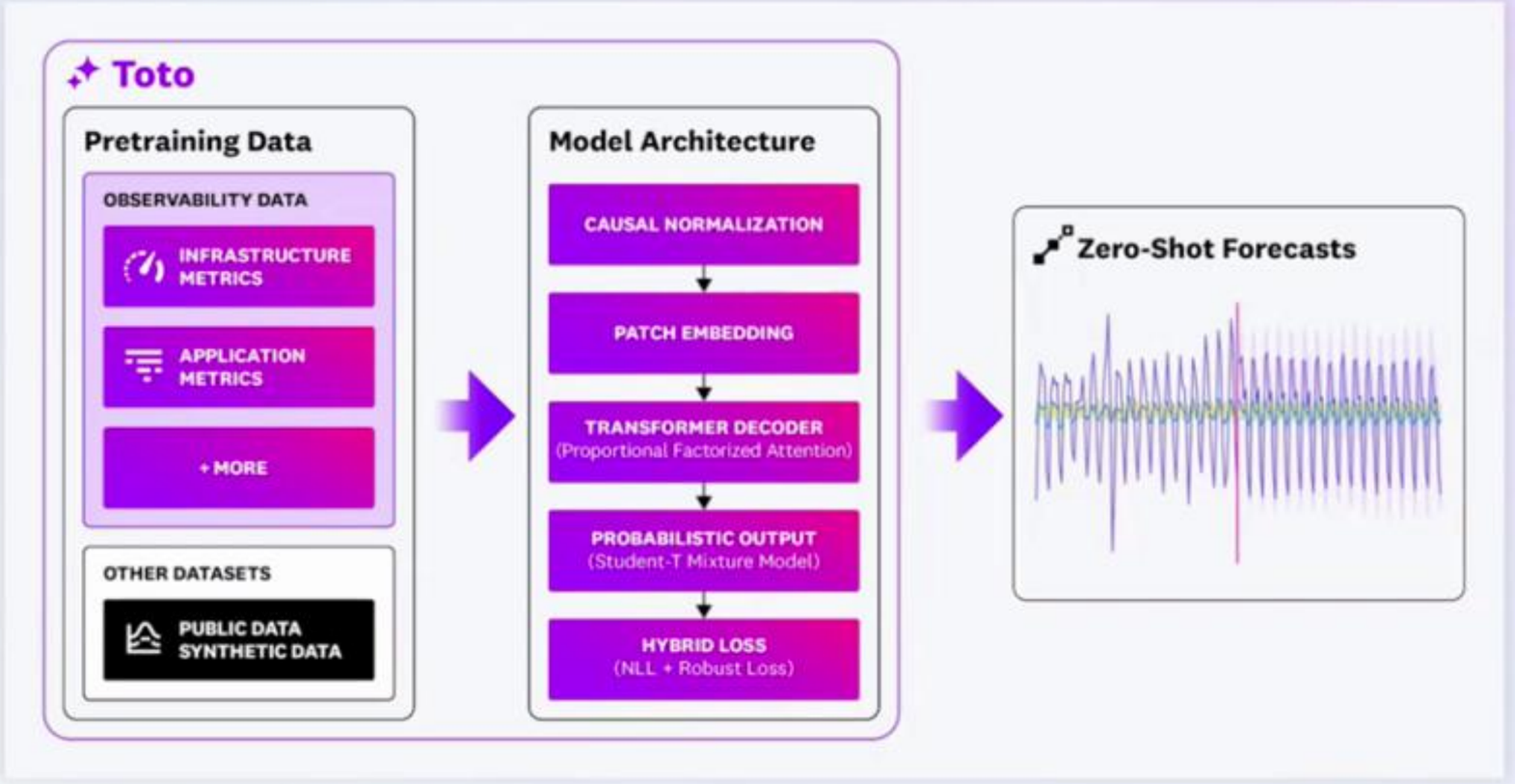
Optimized for Observability

And also SOTA on general-purpose time series forecasting

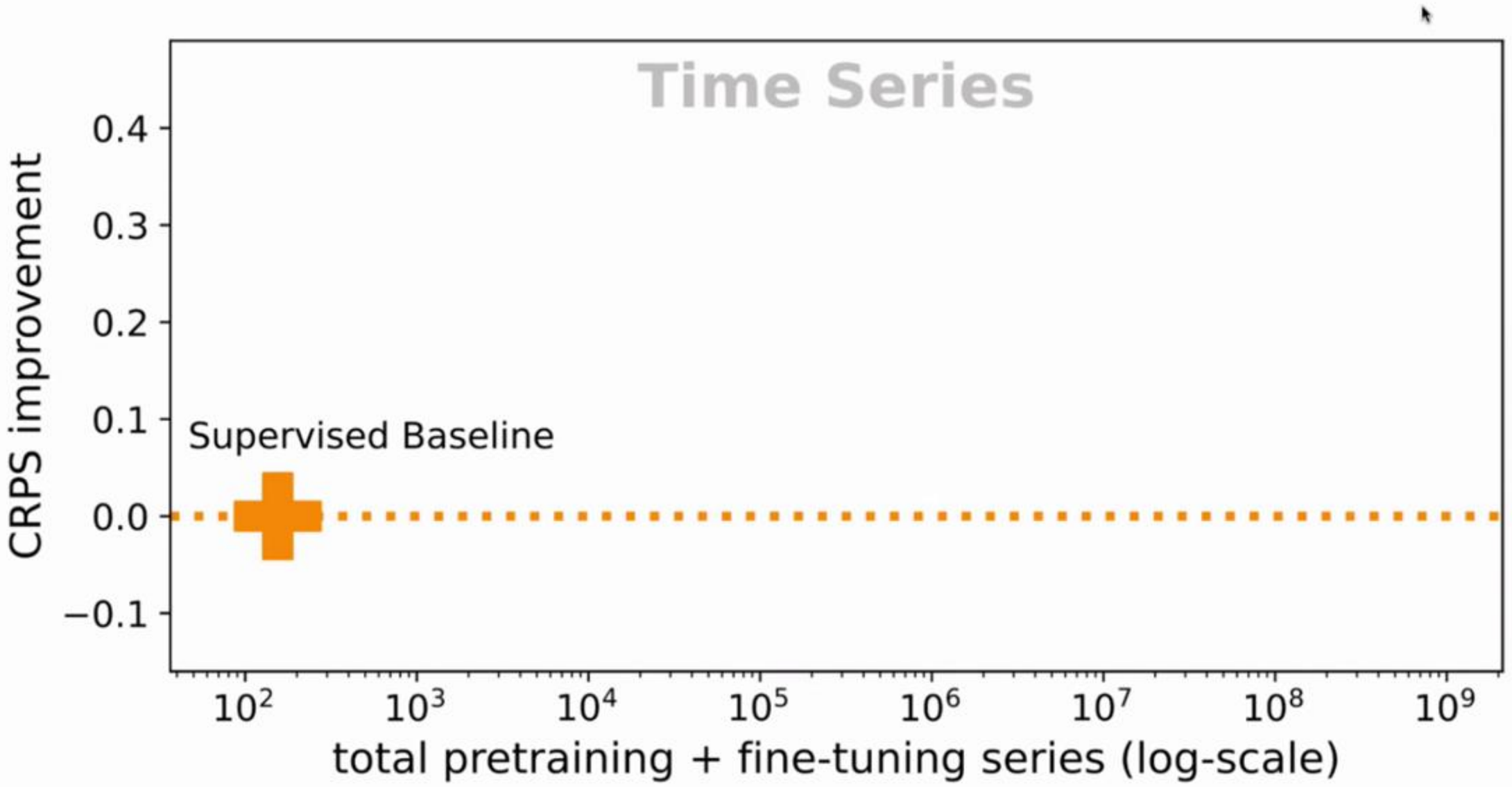
Open Source/Weights

Apache 2.0

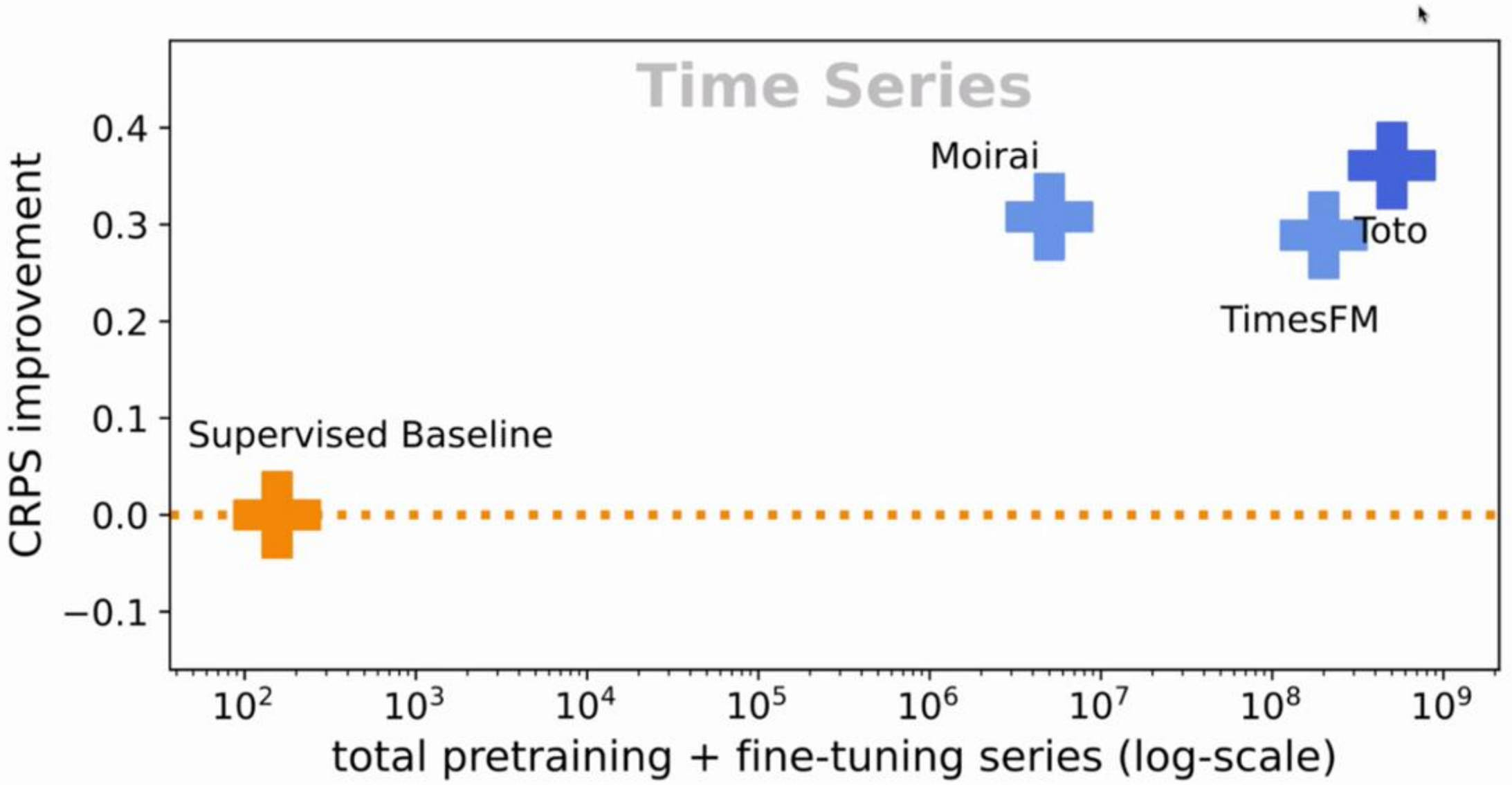
Downloaded >9M times from HF



BOOM Results



BOOM Results



Scaling Toto – From the BERT moment to the GPT-2 Moment

Scaling Toto – From the BERT moment to the GPT-2 Moment

Toto-2.0

DATADOG

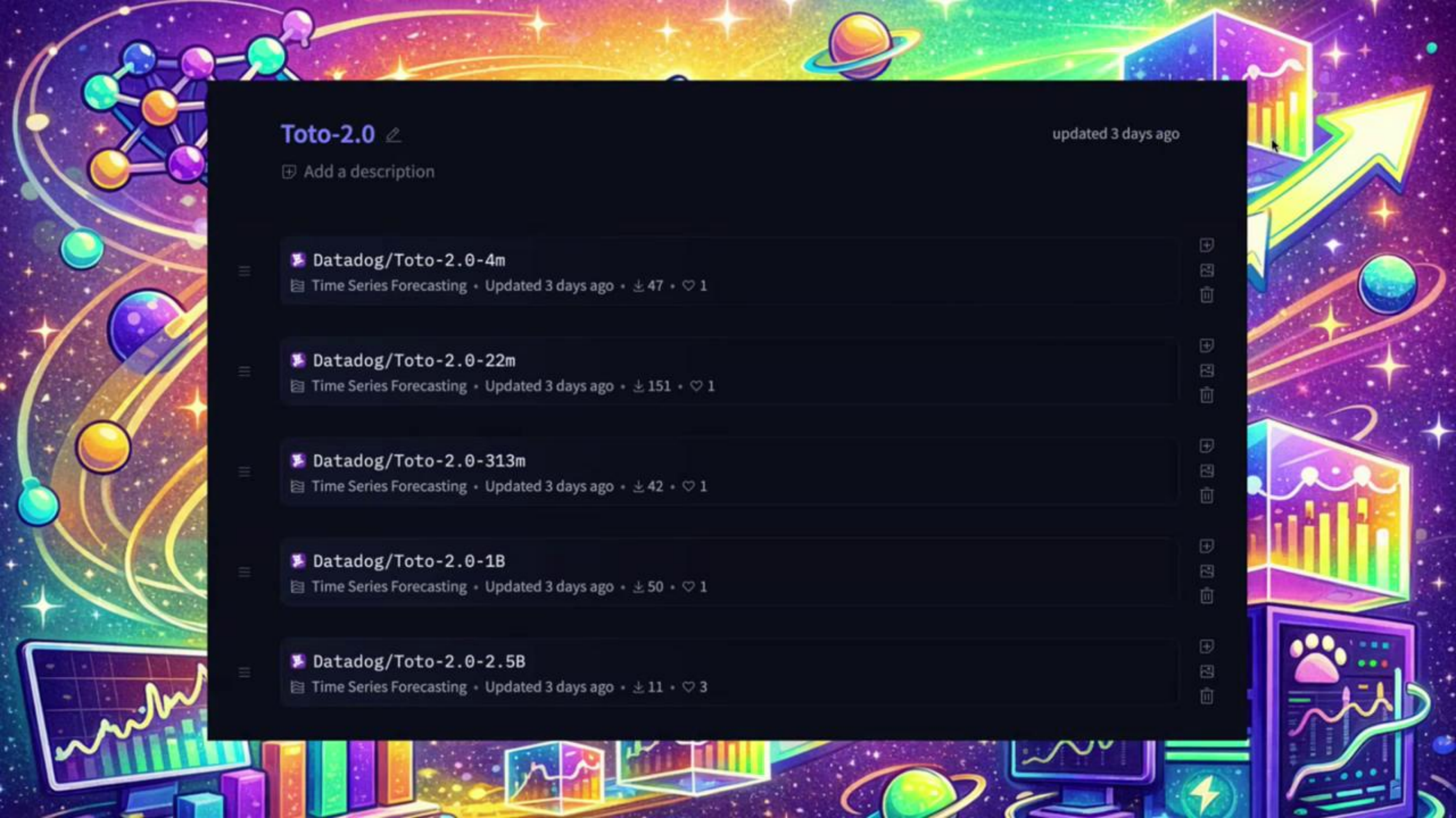

Toto-2.0


Toto-2.0

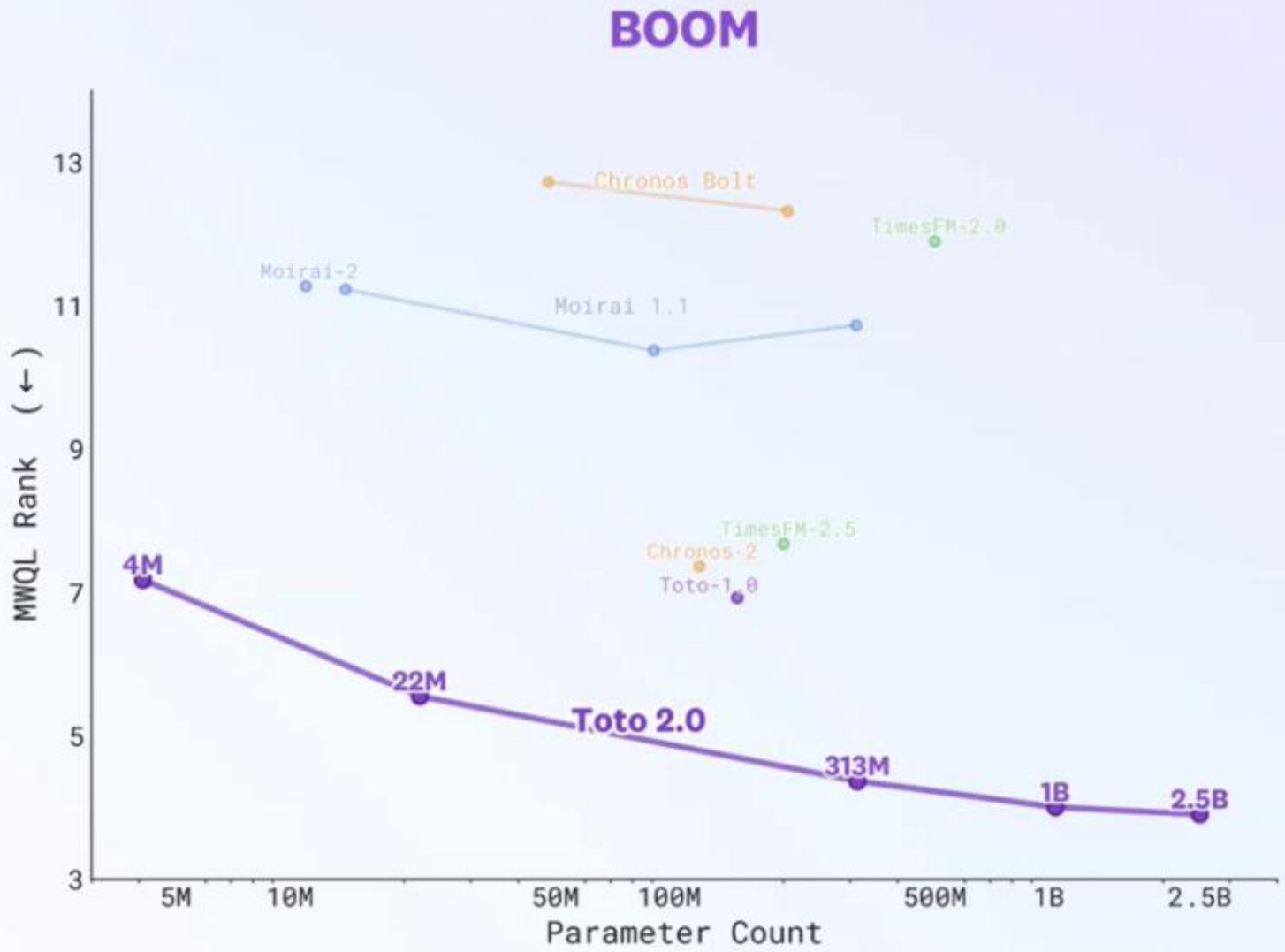

Toto-2.0


Toto-2.0


Toto-2.0


Toto-2.0
updated 3 days ago
Add a description
Datadog/Toto-2.0-4m
Time Series Forecasting • Updated 3 days ago • 47 • 1
Datadog/Toto-2.0-22m
Time Series Forecasting • Updated 3 days ago • 151 • 1
Datadog/Toto-2.0-313m
Time Series Forecasting • Updated 3 days ago • 42 • 1
Datadog/Toto-2.0-1B
Time Series Forecasting • Updated 3 days ago • 50 • 1
Datadog/Toto-2.0-2.5B
Time Series Forecasting • Updated 3 days ago • 11 • 3

tl;dr We can scale to a SOTA model (the GPT-2 Moment)



DATADOG

tl;dr We can scale to a SOTA model (the GPT-2 Moment)

DATADOG


DATADOG

Emaad Khwaja, Chris Lettieri, Gerald Woo

Eden Belouadah, Marc Cenac, Guillaume Jarry, Enguerrand Paquin, Xunyi Zhao, Viktoriya Zhukova

Othmane Abou-Amal, Ameet Talwalkar, David Asker

DATADOG

Emaad Khwaja, Chris Lettieri, Gerald Woo

Eden Belouadah, Marc Cenac, Guillaume Jarry, Enguerrand Paquin, Xunyi Zhao, Viktoriya Zhukova

Othmane Abou-Amal, Ameet Talwalkar, David Asker

What changed from Toto 1.0?

Architecture

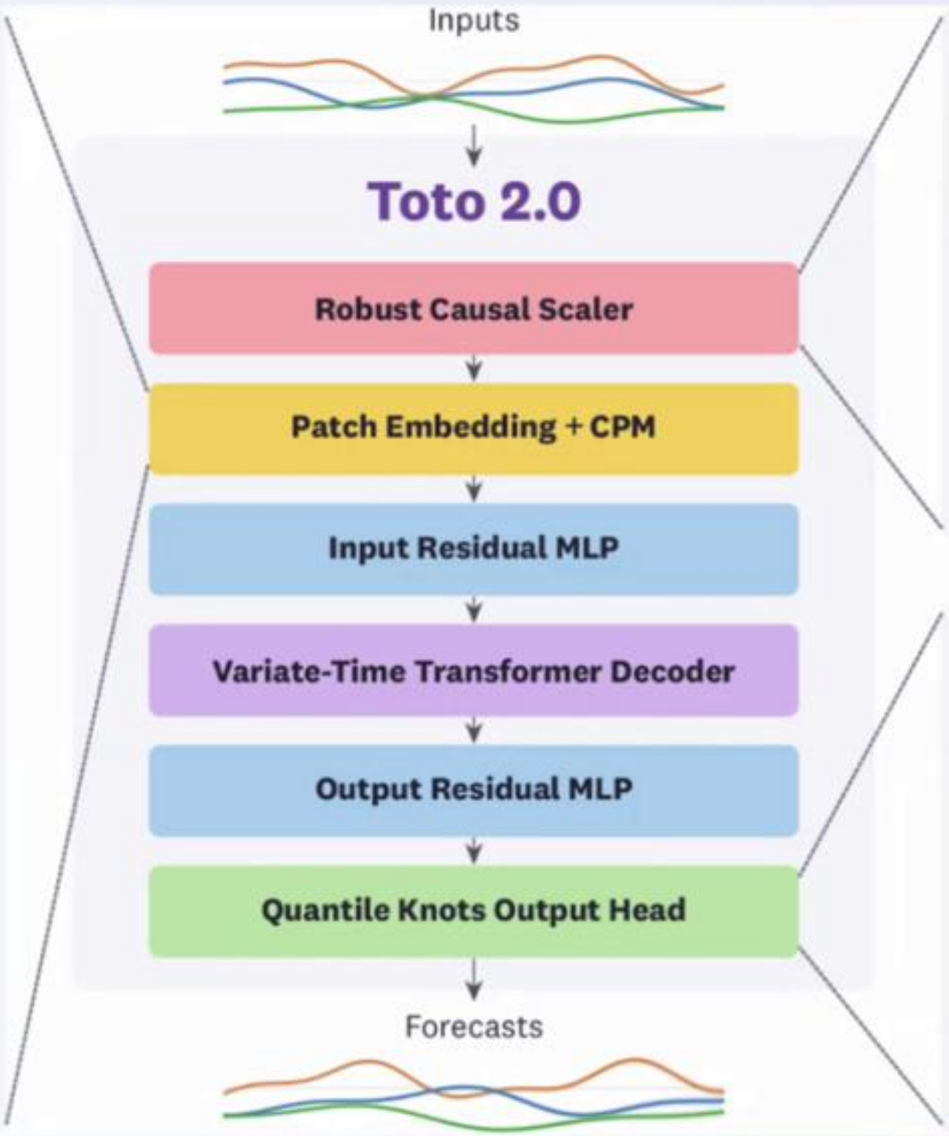
Contiguous Patch Masking (CPM)

Training

target shifted by one patch

Inference

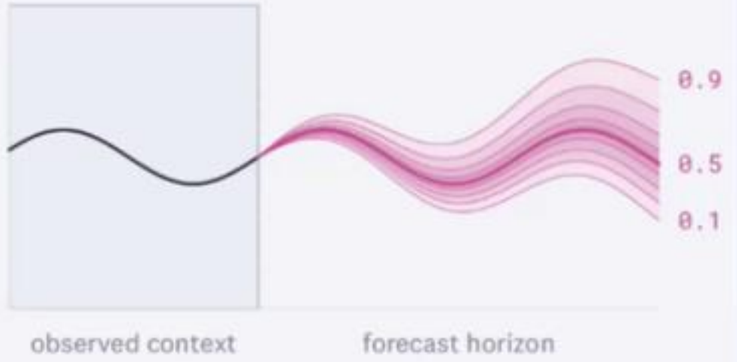
full horizon in one forward pass



Robust Causal Scaler

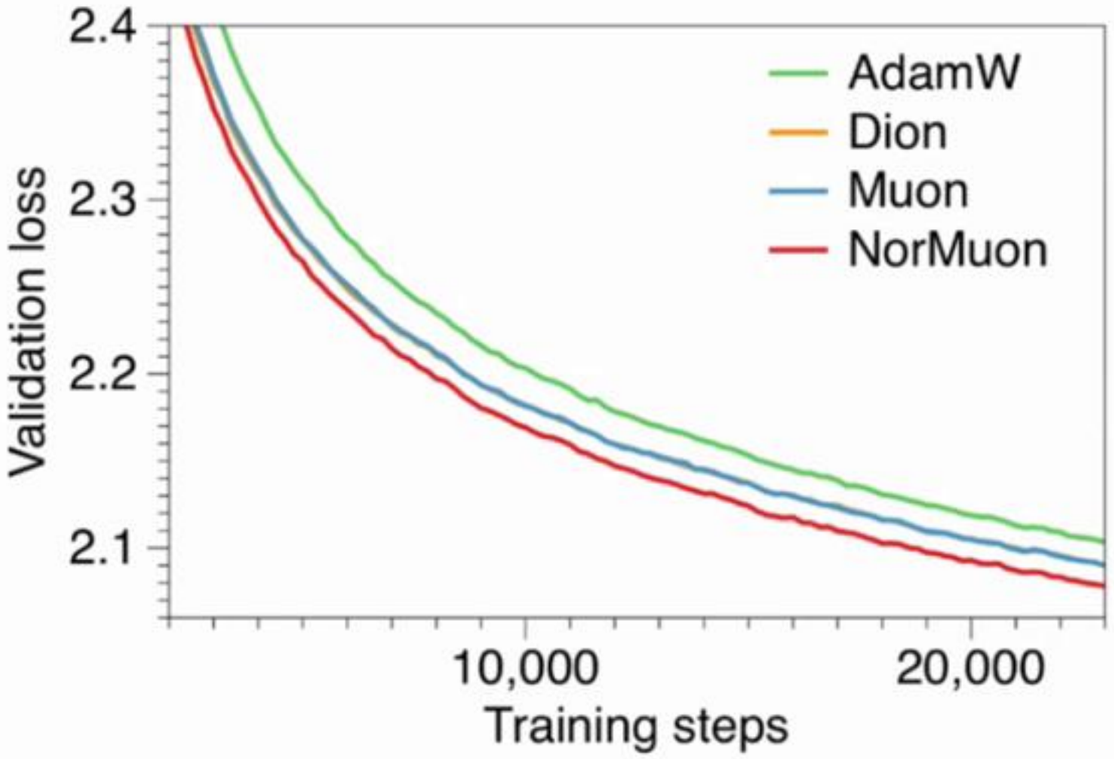
$$z_t = asinh((x_t - \mu_t)/\sigma_t)$$

Quantile Knots Head

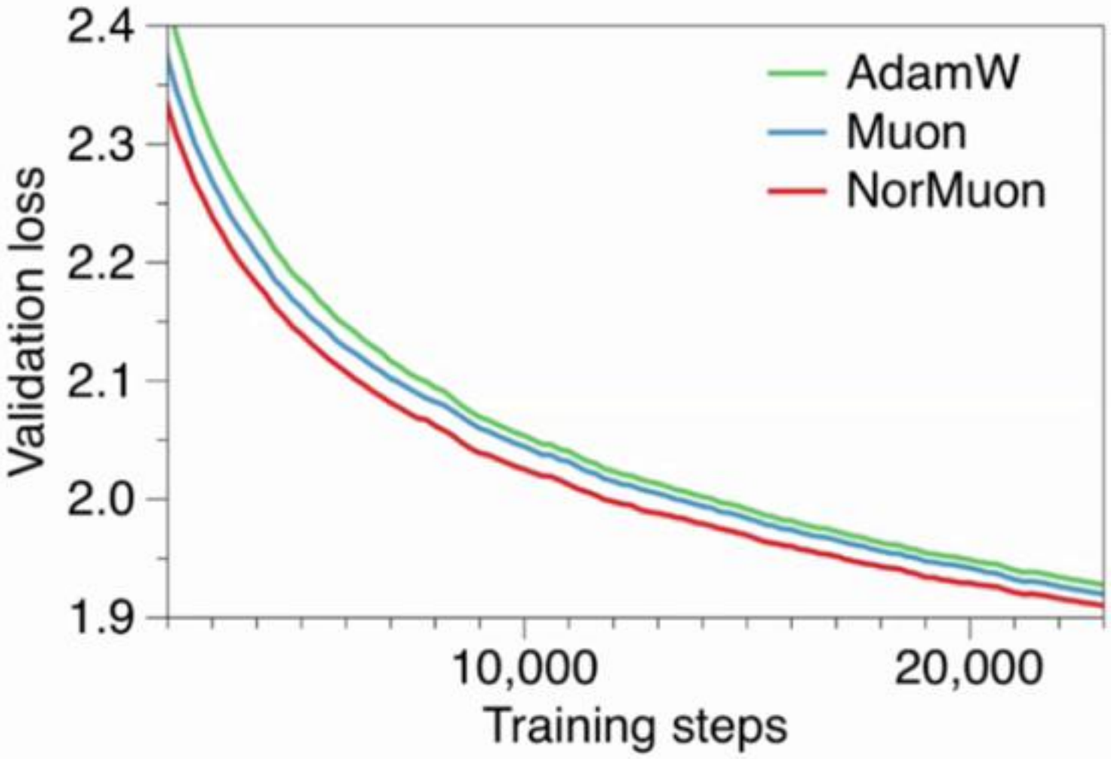


DATADOG

Optimizer: Normuon + AdamW



(a) Pretraining results of 1.1B model.



(b) Pretraining results of 5.4B model.

DATADOG

Beating GPT-2 for <<$100: the nanochat journey #481

karpathy announced in Announcements

karpathy on Jan 31 Maintainer

edited

When OpenAI released GPT-2 in February 2019, training the largest model (1.5B parameters) required serious compute:

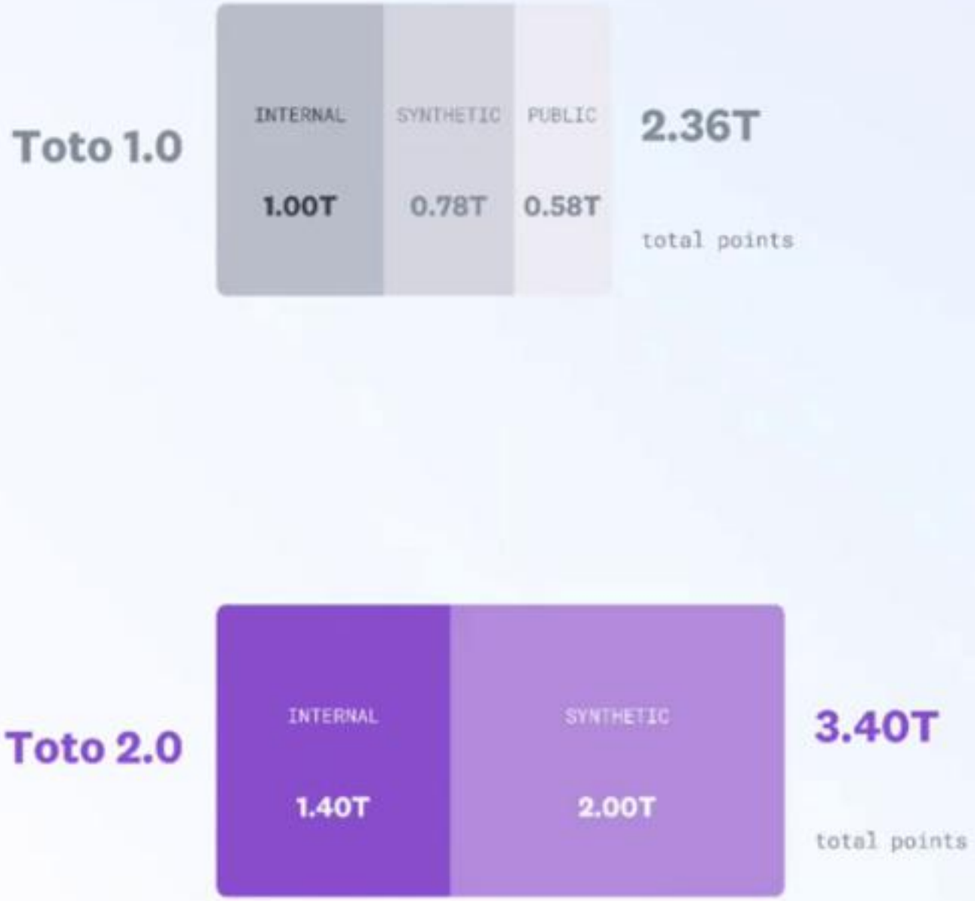
- **Hardware:** 32 TPU v3 chips (256 TPU v3 cores, 8 cores per chip)
- **Training time:** "A bit over a week" (~168 hours)
- **Cloud cost:** At $8/hour per TPU v3, that's `32 × 168 × $8 =` **$43,000**

Sources: Reddit thread from 2019, HuggingFace model card.

Beating GPT-2 for <$100 from scratch has been a bit of an odd obsession for me but finally here we are. Seven years later, we can beat GPT-2's performance in nanochat ~1000 lines of code running on a single 8XH100 GPU node for ~3 hours. At ~$24/hour for an 8×H100 node, that's **$73**, i.e. **~600× cost reduction**. That is, each year the cost to train GPT-2 is falling to approximately 40% of the previous year. (I think this is an understimate and that further improvements are still quite possible). The gains come from everywhere: better hardware (H100 vs TPU v3), better software (Flash Attention 3, torch.compile), better algorithms (Muon optimizer, architectural improvements), and better data (FineWeb-edu).

Toto 2.0 Training Data

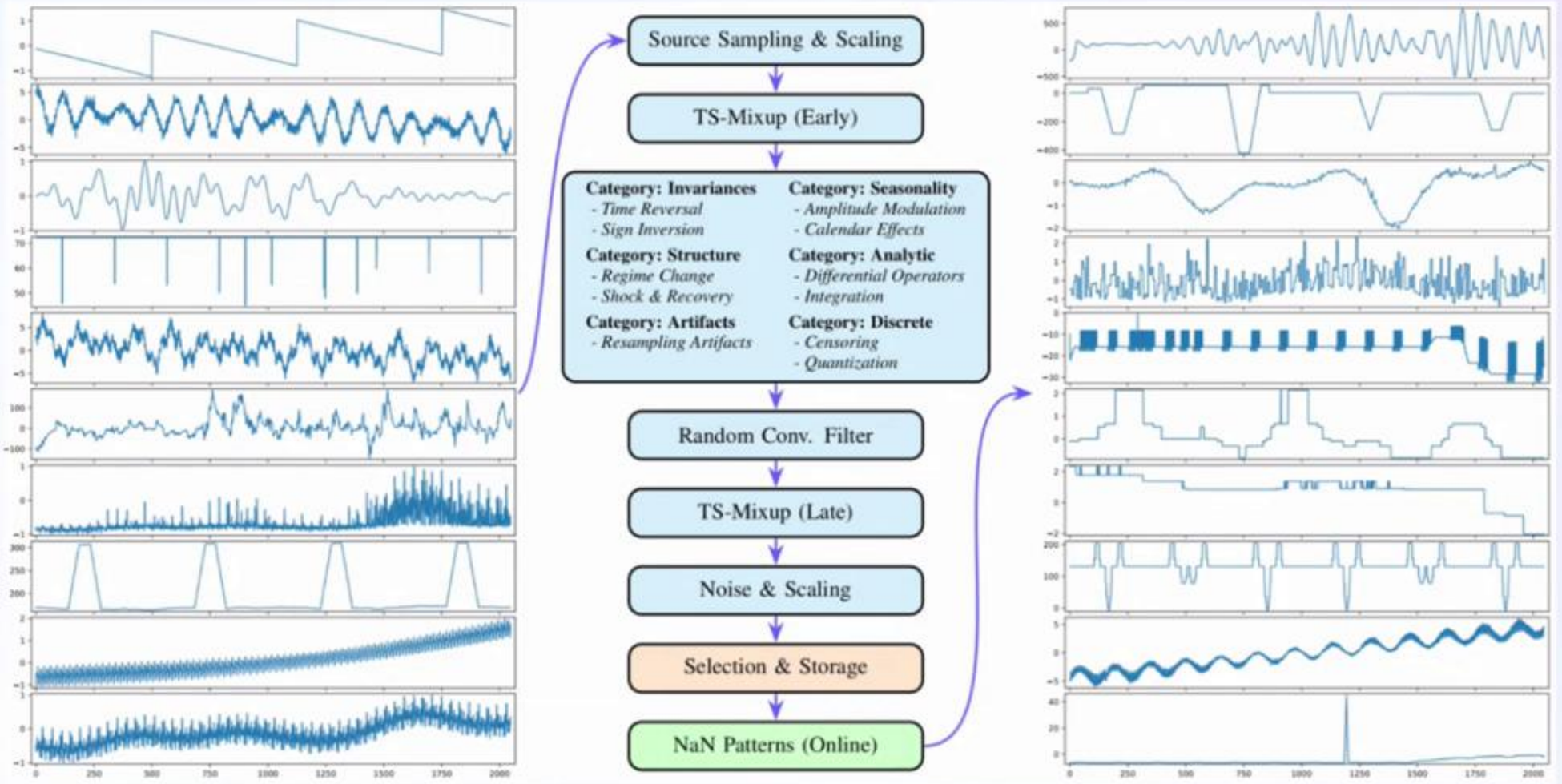
Composition



Interval

DATADOG

TempoPFN Synthetic Data



DATADOG

Scaling Efficiently


u-μP
Proxy Model
Target Models


u-μP
Proxy Model
Target Models

Searching for Model Parameters with REx

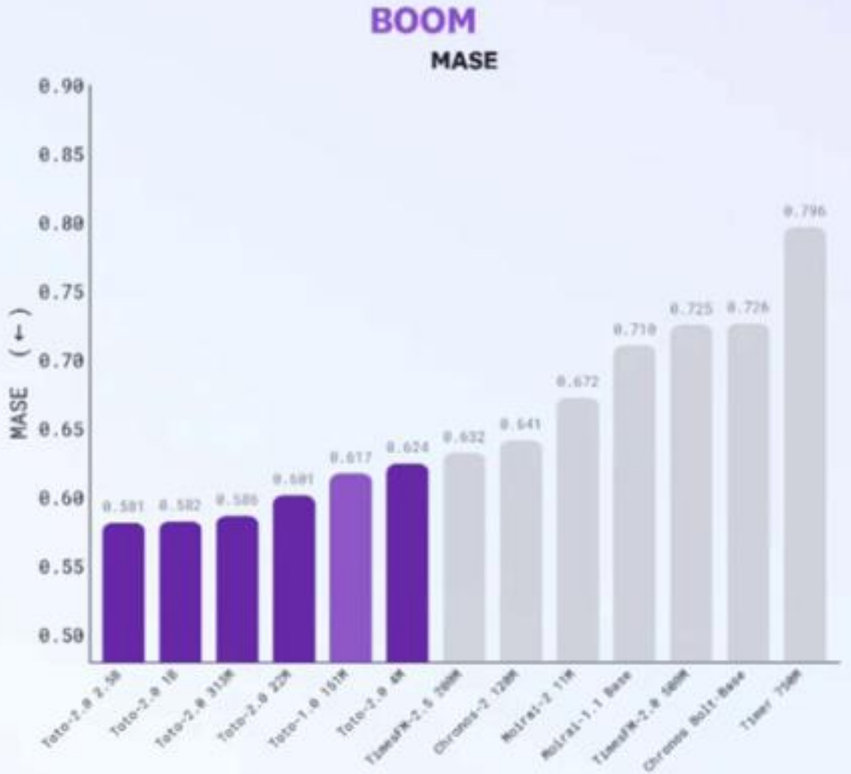
Round 1: Architecture Choices → Round 2: Data Selection → Round 3: Optimizer → Round 4: Decay Schedule

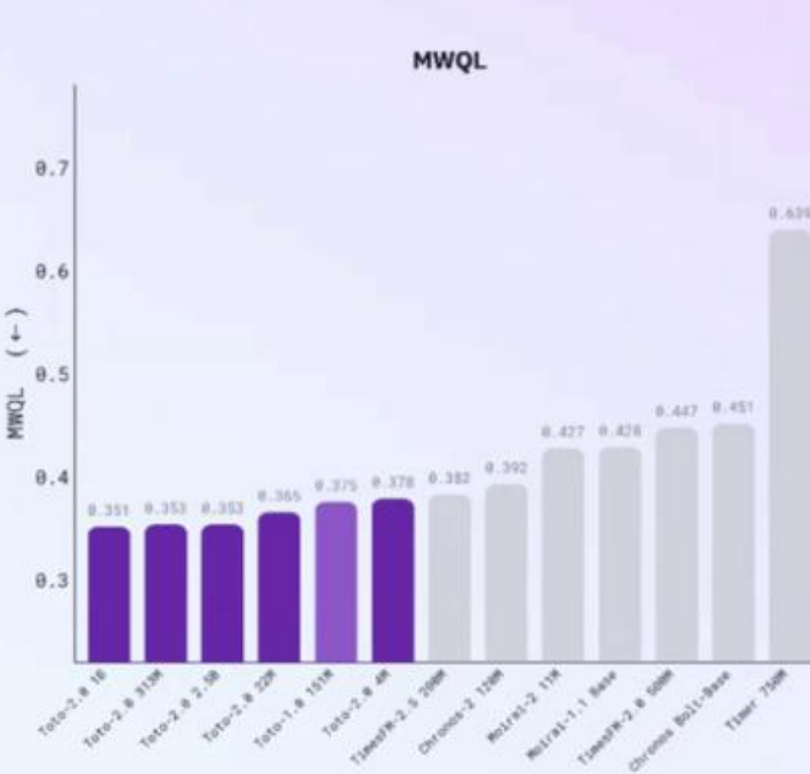
Zero shot scale up to 2.5B!

DATADOG

Results

BOOM



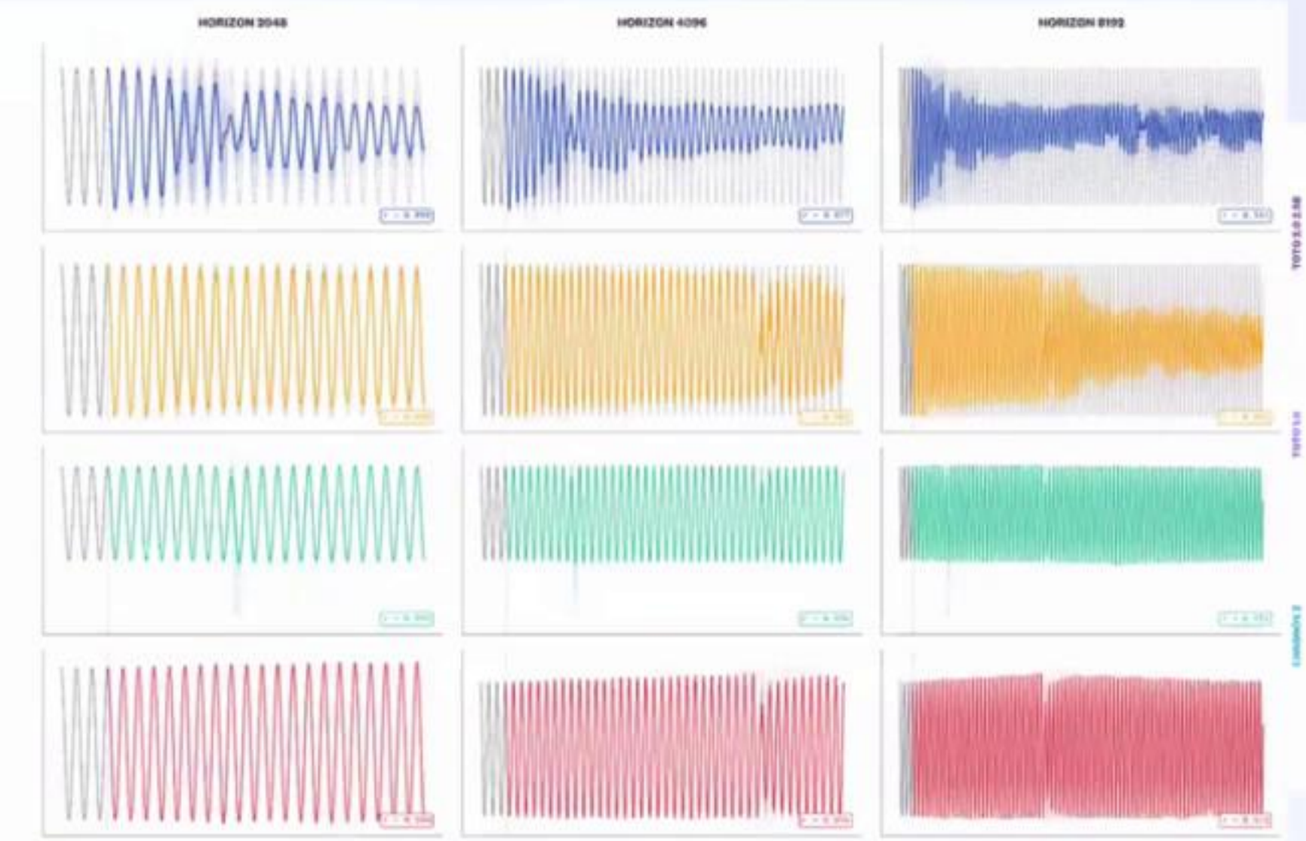


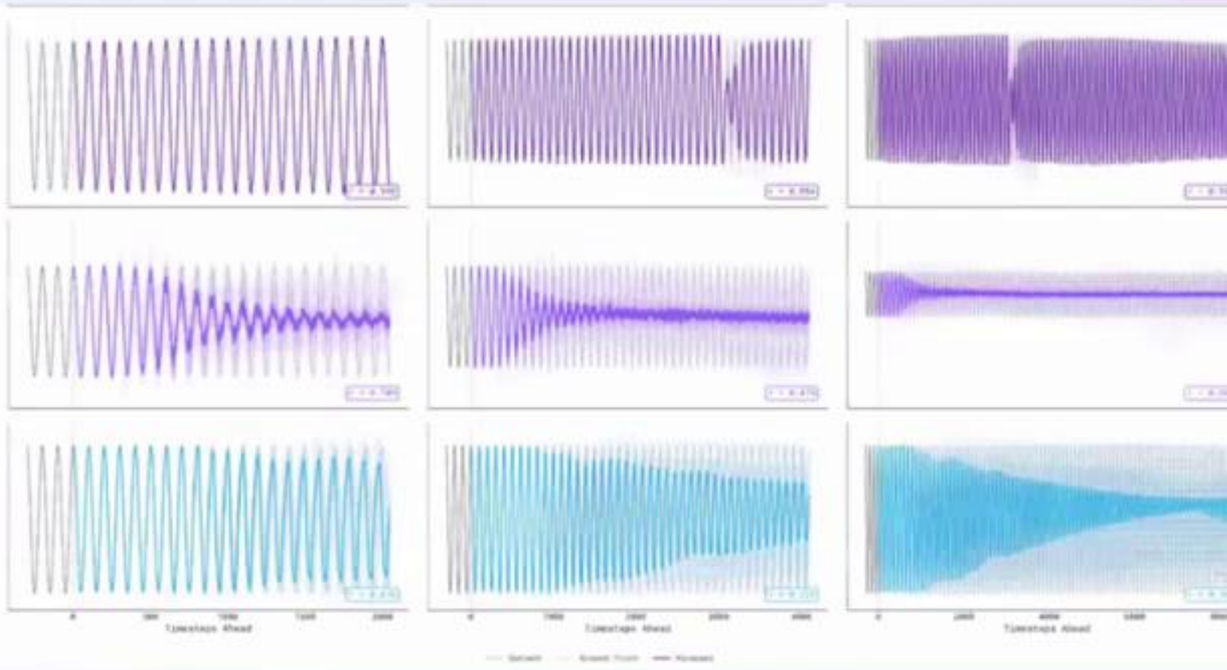
DATADOG

GIFT-Eval (update in progress)

T	model	Organization	Test Leak.	Replication Cod	MASE	MASE_Rank	CRPS	CRPS_Rank
🟢	Toto-2.0-2.5B (code)	Datadog	No	Yes	0.698	19.041	0.481	18.031
🟢	Toto-2.0-1B (code)	Datadog	No	Yes	0.700	19.701	0.48	18.546
🟢	Chronos-2 (code)	AWS	No	Yes	0.698	19.928	0.485	20.093
🟢	Toto-2.0-313m (code)	Datadog	No	Yes	0.701	21.474	0.486	19.918
🟢	Timer-S1 (code)	Tsinghua & Bytel	No	Yes	0.693	24.052	0.485	24.01
🔴	PatchTST-FM-r1 (code)	IBM TSFM & Rens	No	Yes	0.707	24.175	0.483	20.68
🔴	FlowState-r1.1 (code)	IBM TSFM	No	Yes	0.701	24.227	0.487	22.918
🔴	TimesFM-2.5 (code)	Google Research	No	Yes	0.705	24.567	0.49	23.32
🔴	TiRex (code)	NXAI	No	Yes	0.716	26.742	0.488	21.412
🔴	Granite-PatchTST-FM-r1	IBM TSFM & Rens	No	Yes	0.717	27.866	0.488	23.124
🟢	Toto-2.0-22m (code)	Datadog	No	Yes	0.719	28.258	0.497	24.526
🔴	Chronos-2-Synth (code)	AWS	No	Yes	0.720	29.01	0.496	27.918

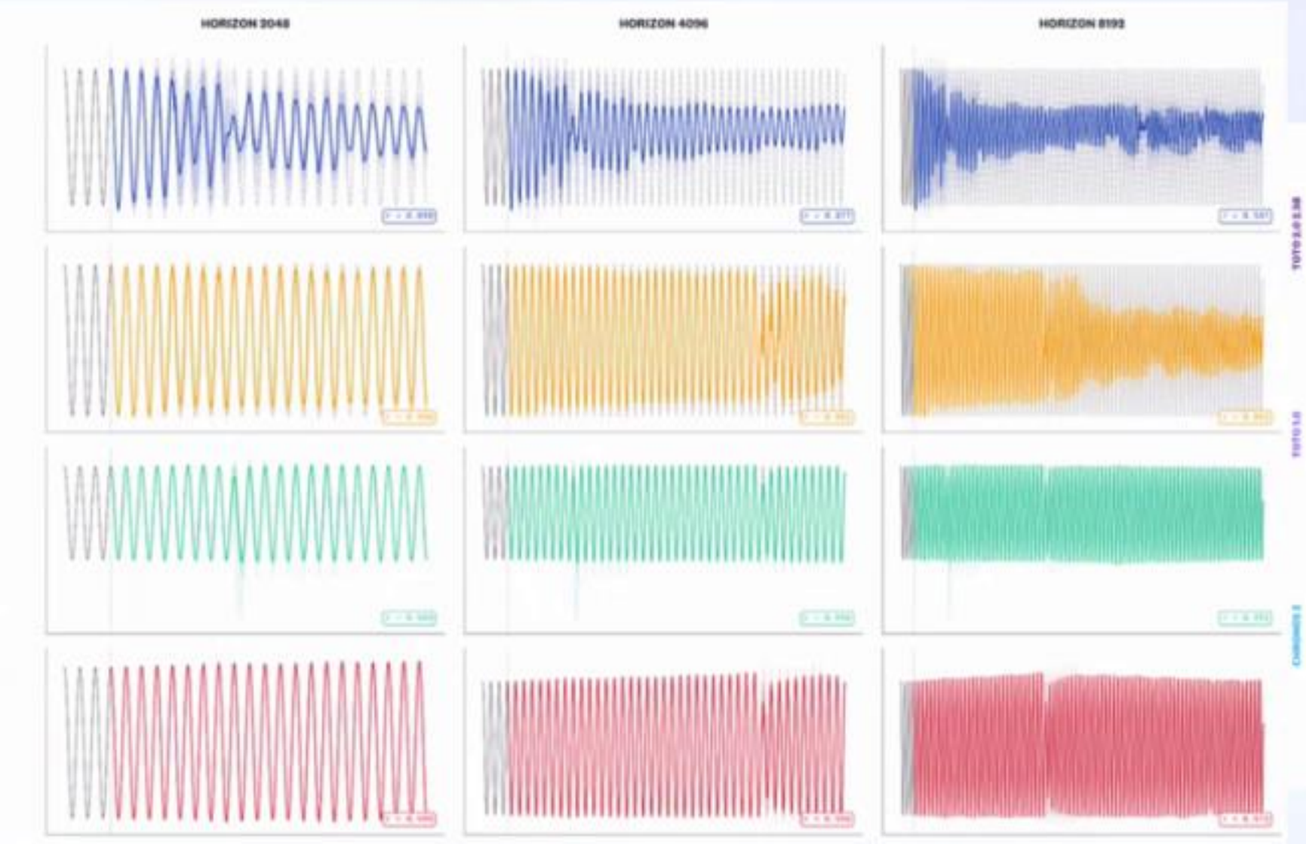
Discovery: Long Horizon Stability Scales with Size

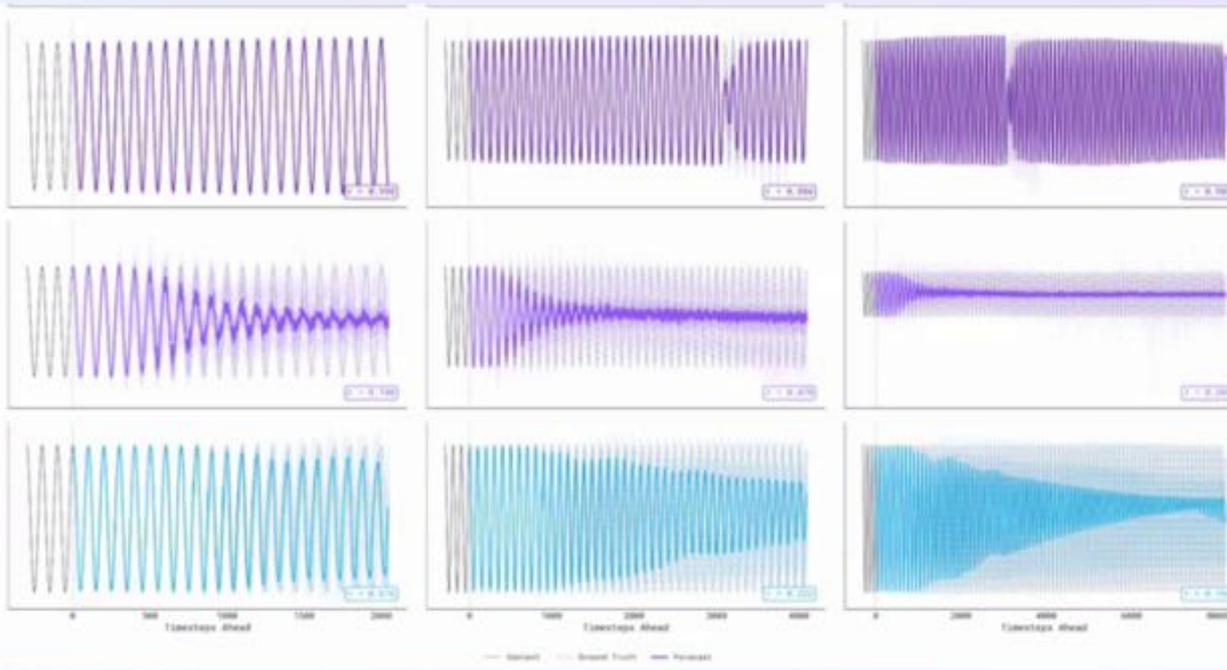




DATADOG

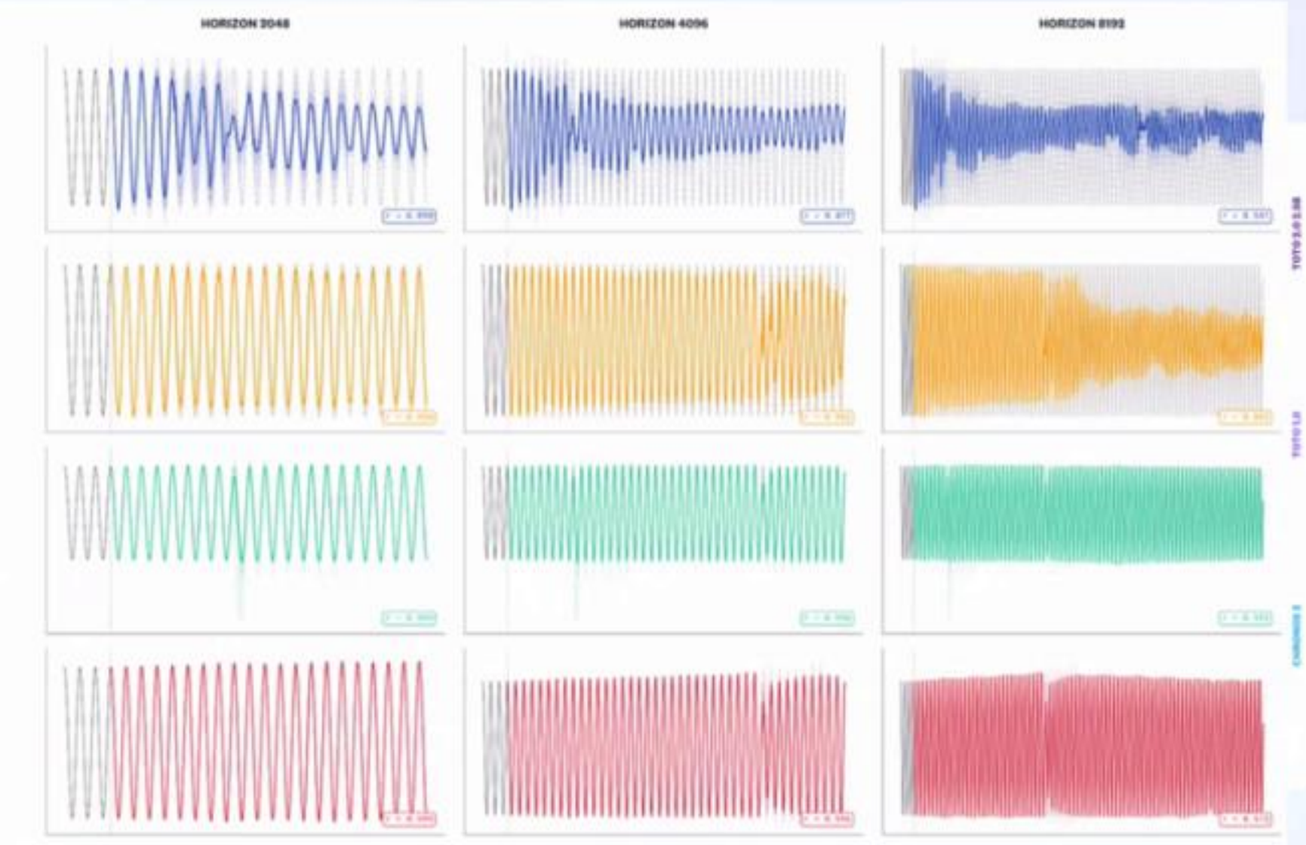
Discovery: Long Horizon Stability Scales with Size

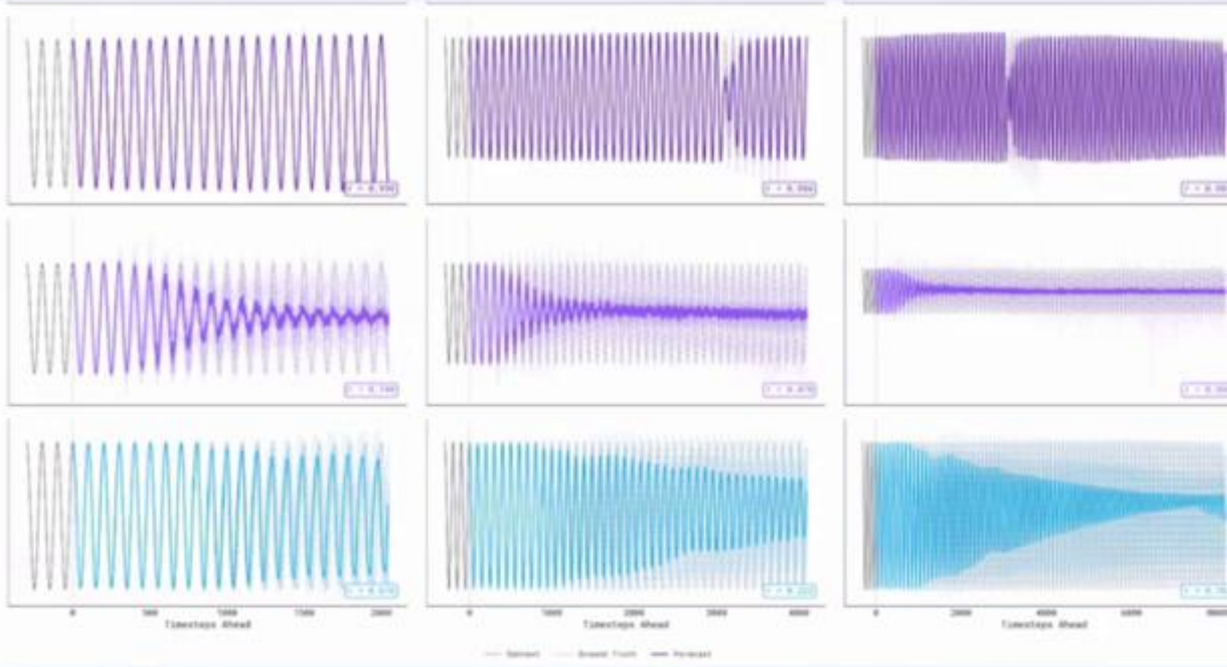




DATADOG

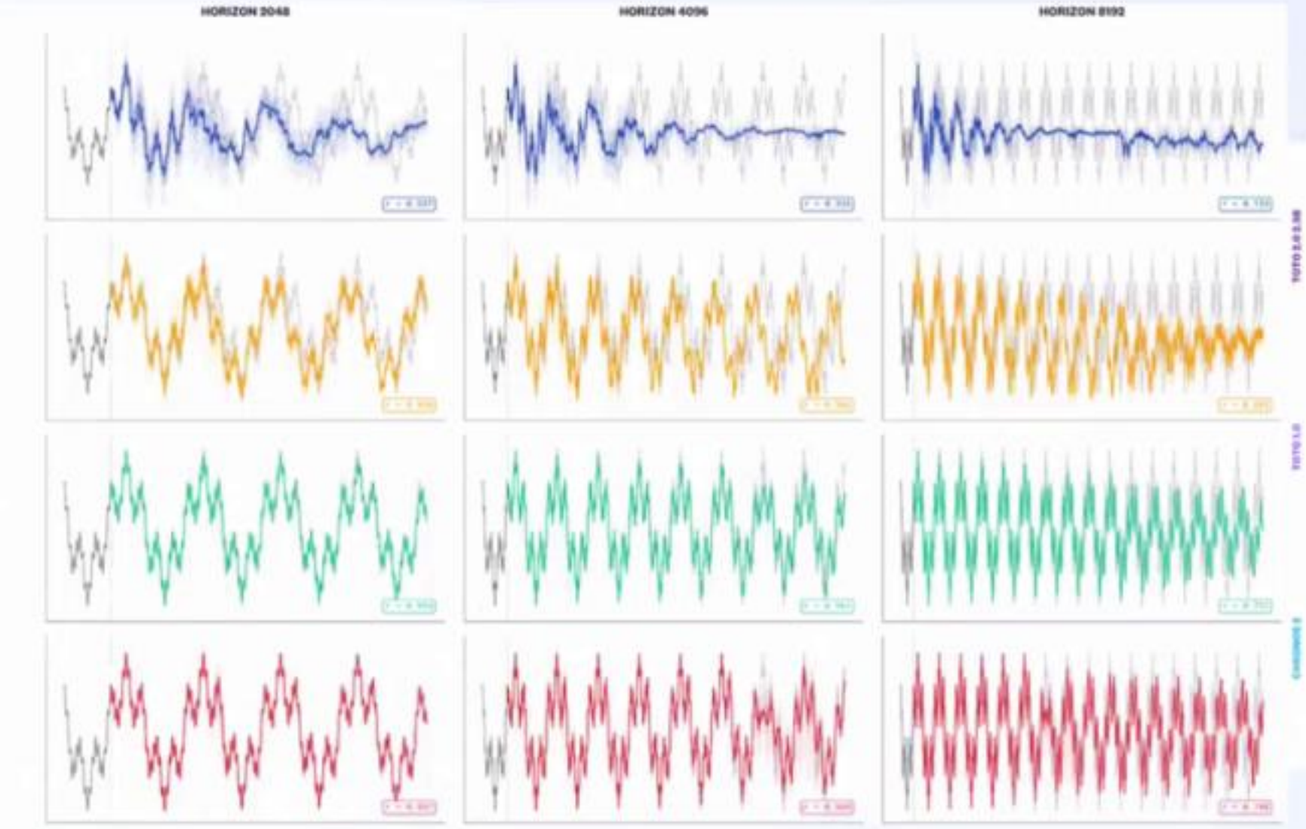
Discovery: Long Horizon Stability Scales with Size

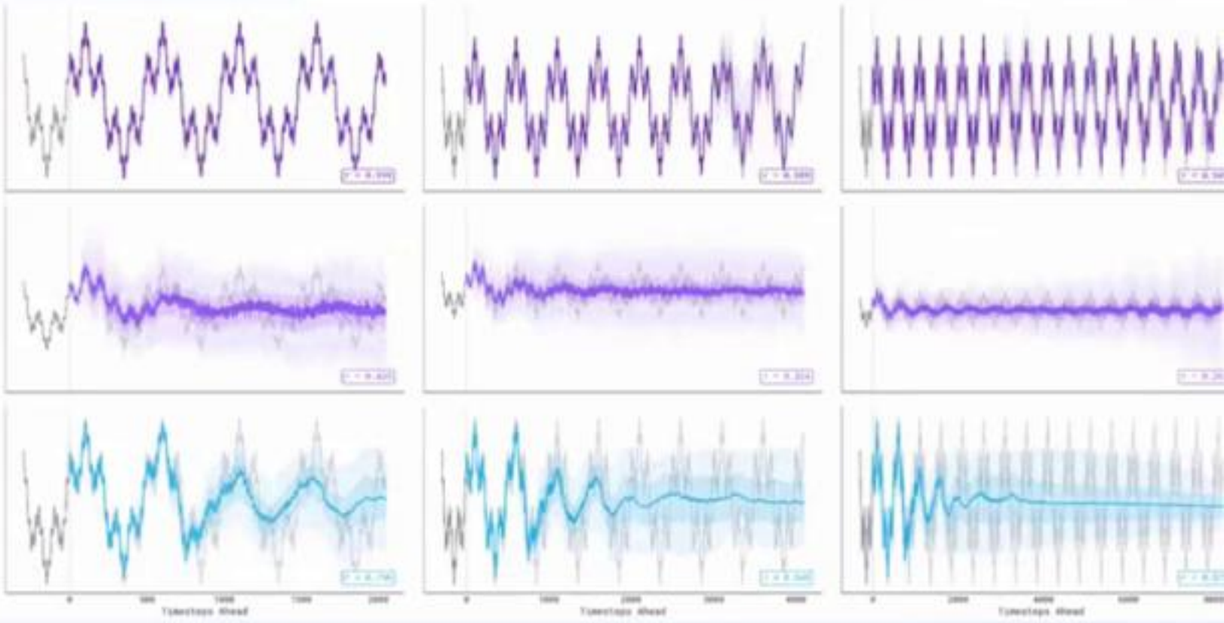




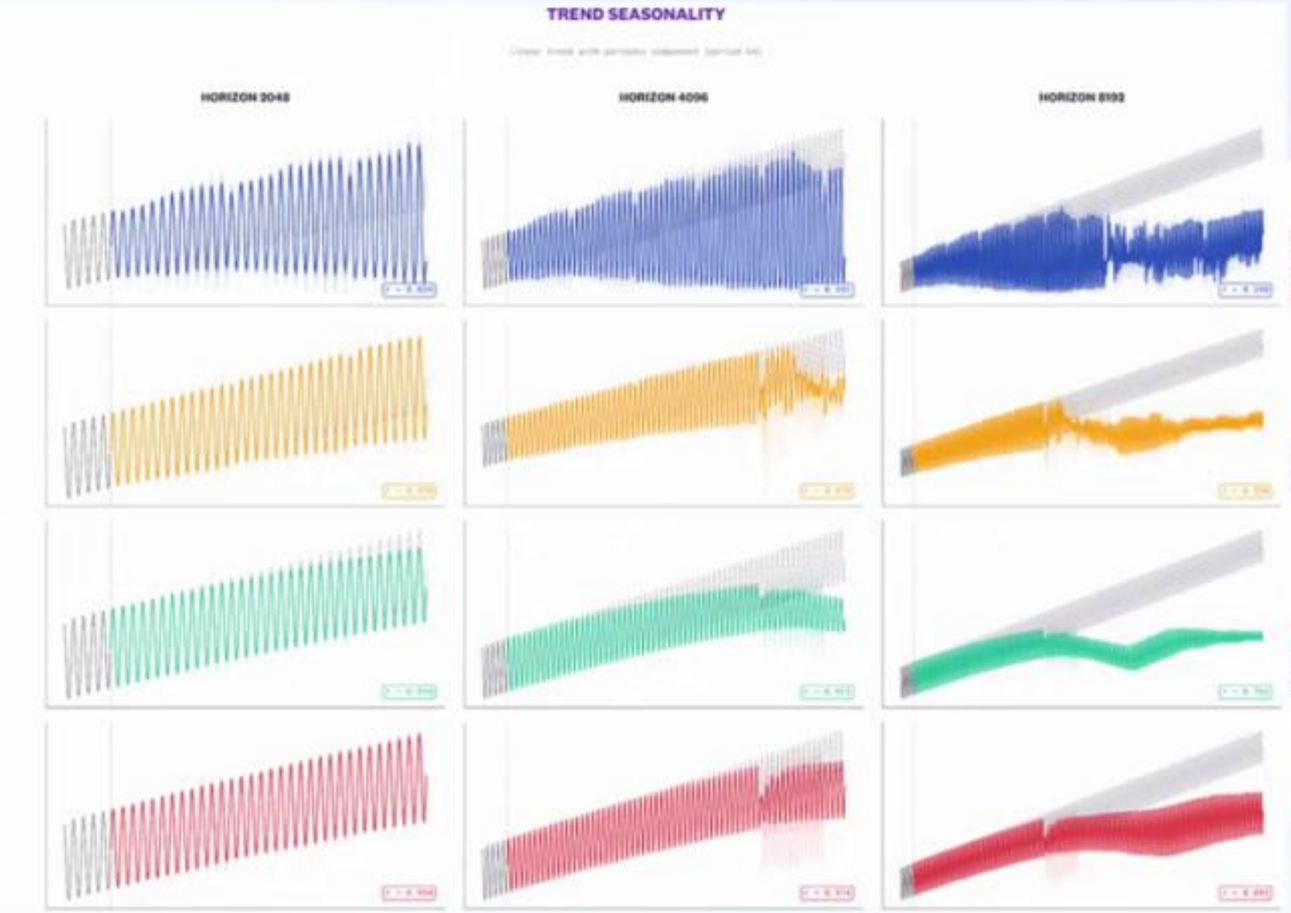
DATADOG

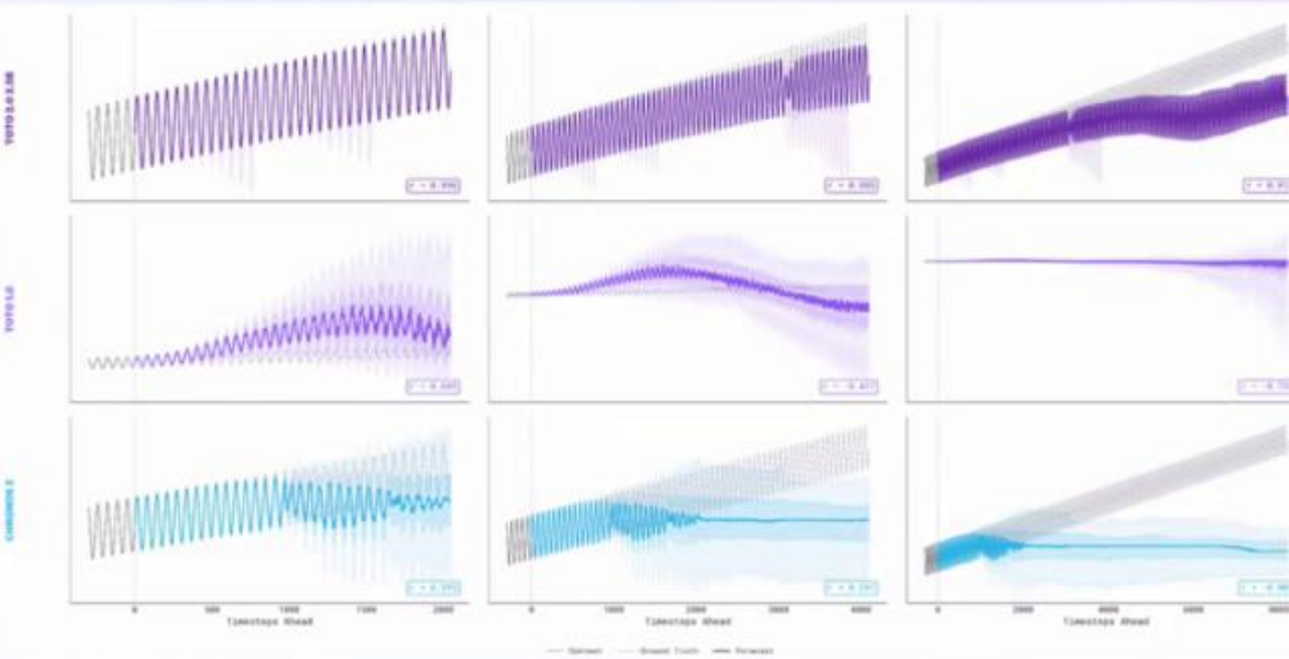
Discovery: Long Horizon Stability Scales with Size





Discovery: Long Horizon Stability Scales with Size





From Time Series Foundation Models to Multi Modal “World Models”

Multi Modality & World Models

ARFBench: TSQA Benchmark based of internal incidents

ARFBench: Benchmarking Time Series Question Answering Ability for Software Incident Response

Stephan Xie[1,2], Ben Cohen[2], Mononito Goswami[3], Junhong Shen[1],
Emaad Khwaja[2], Chenghao Liu[2], David Asker[2], Othmane Abou-Amal[2], Ameet Talwalkar[1,2]

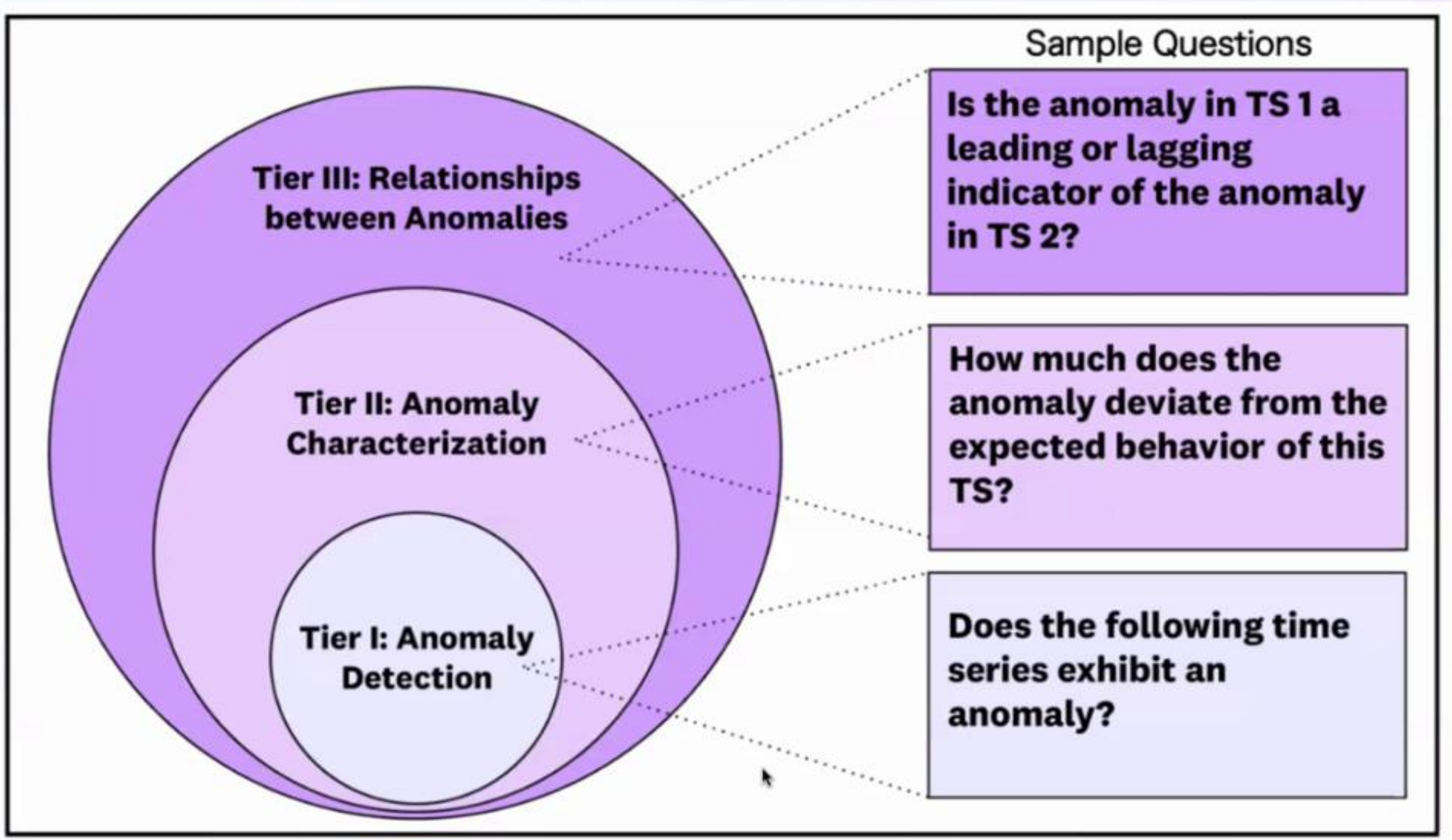
[1]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA
[2]Datadog AI Research, New York, NY, USA
[3]Amazon AI Research, Seattle, WA, USA

{stephan.xie, ameet.talwalkar}@datadoghq.com

DATADOG

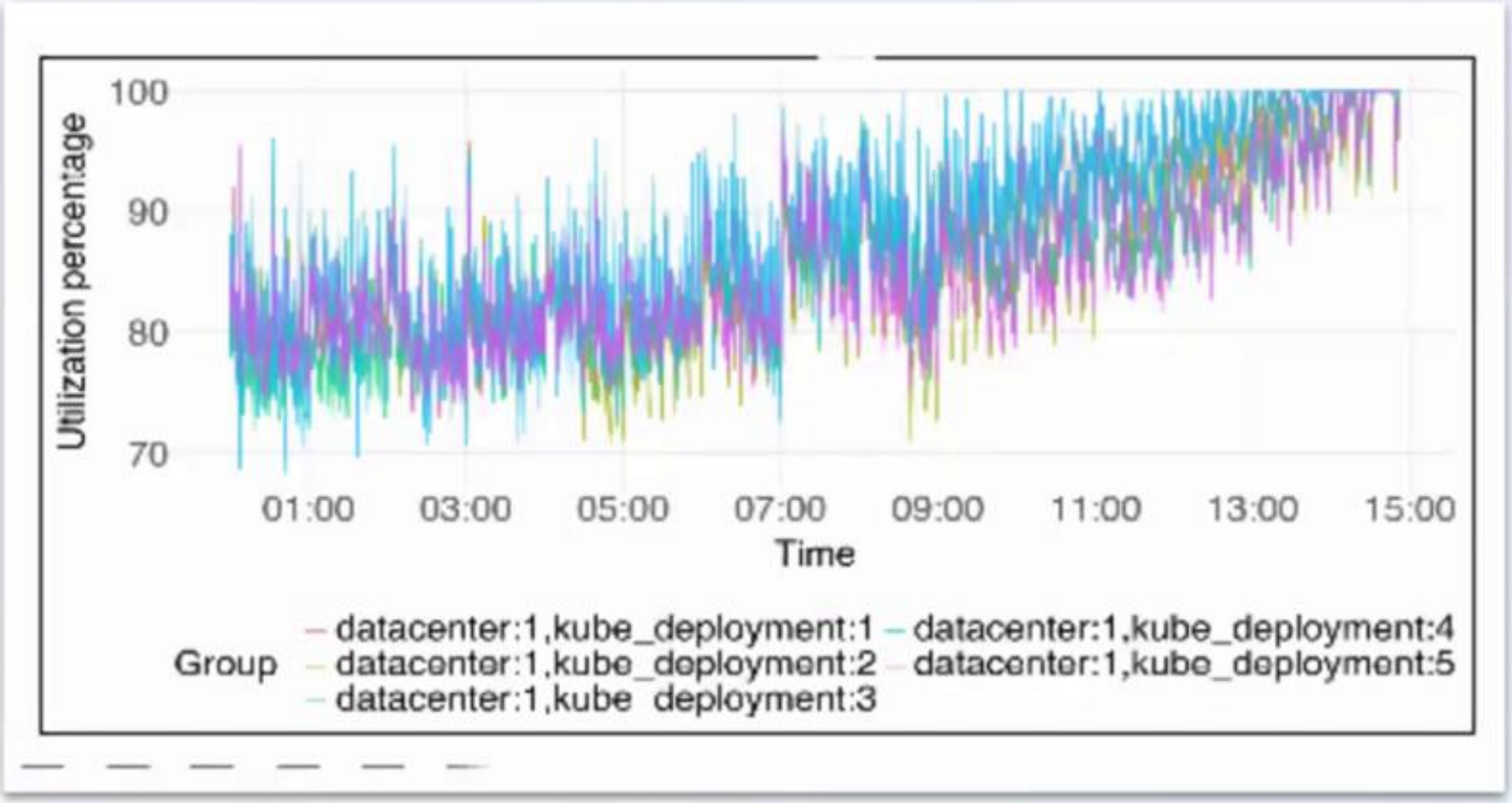
ARFBench: TSQA Benchmark based of internal incidents

ARFBench: TSQA Benchmark based of internal incidents



DATADOG

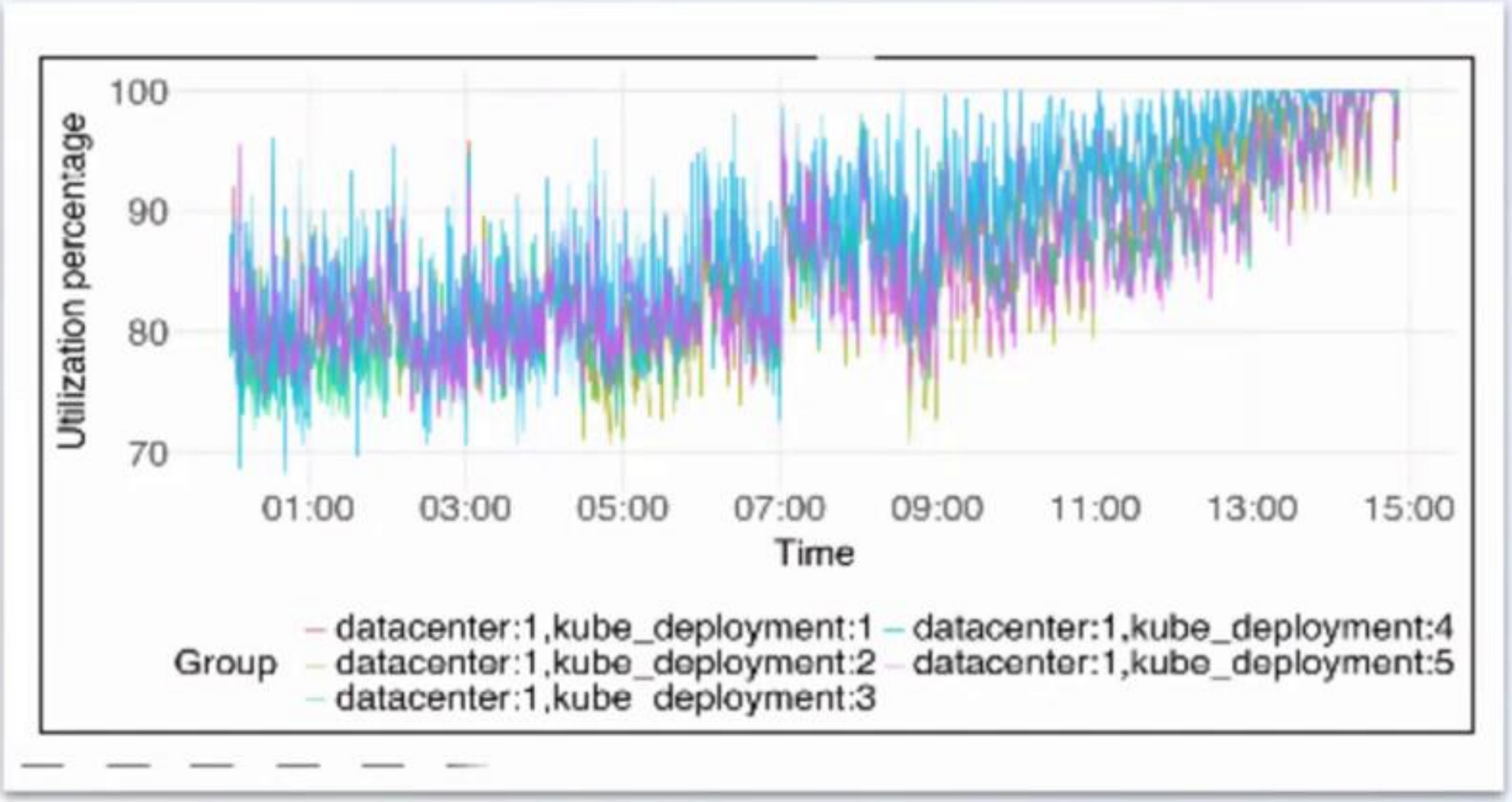
Time series question answering (TSQA)



Avg utilization % for a metadata queue, grouped by k8s deployment and datacenter

DATADOG

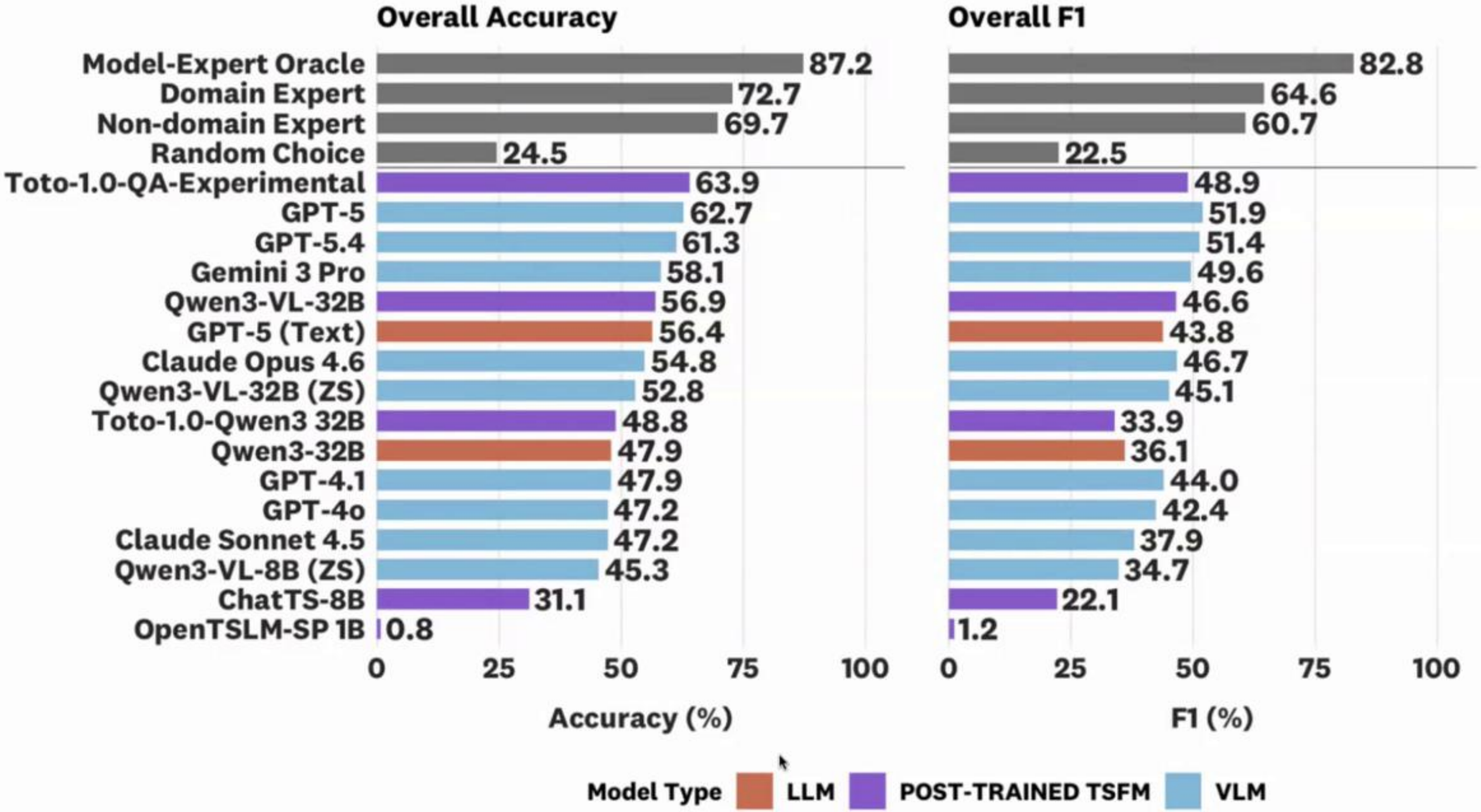
Time series question answering (TSQA)



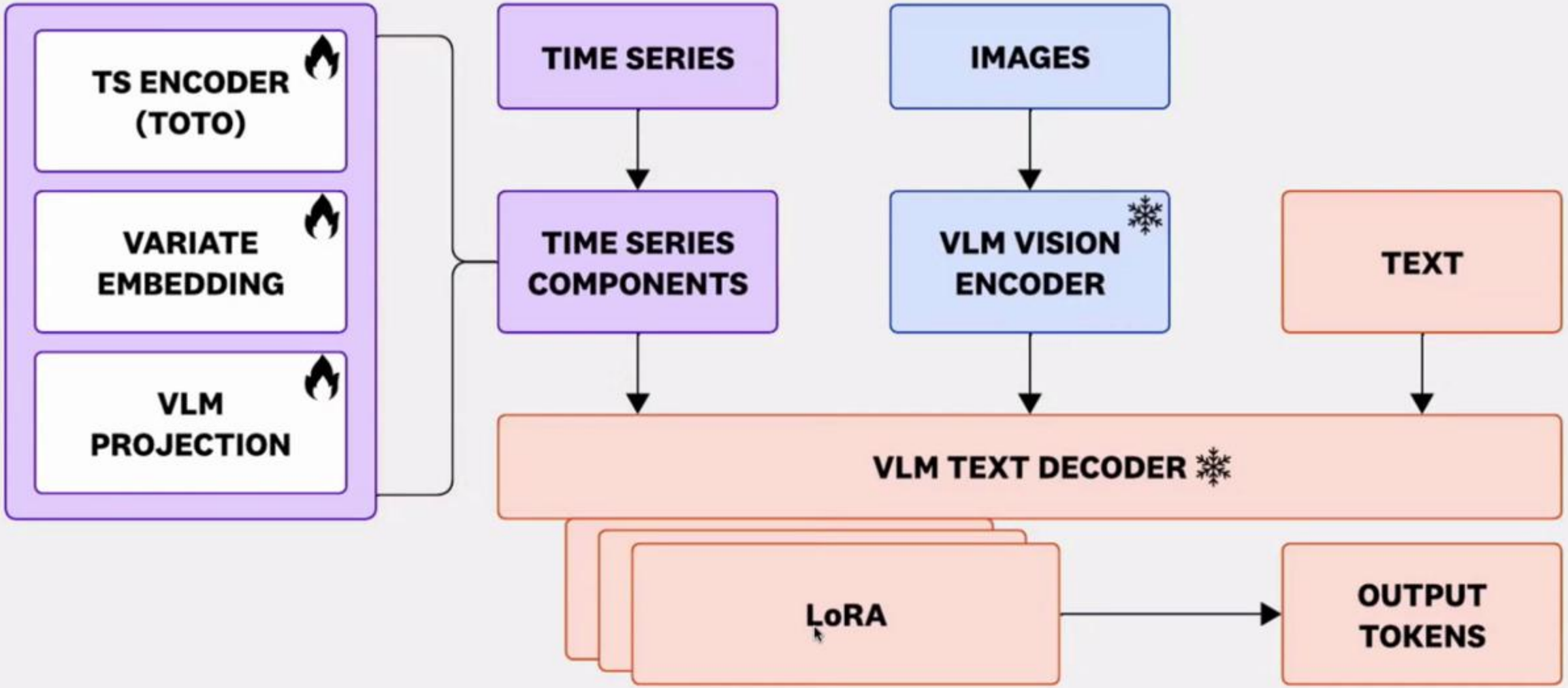
Is there an anomaly in this time series?

Avg utilization % for a metadata queue, grouped by k8s deployment and datacenter

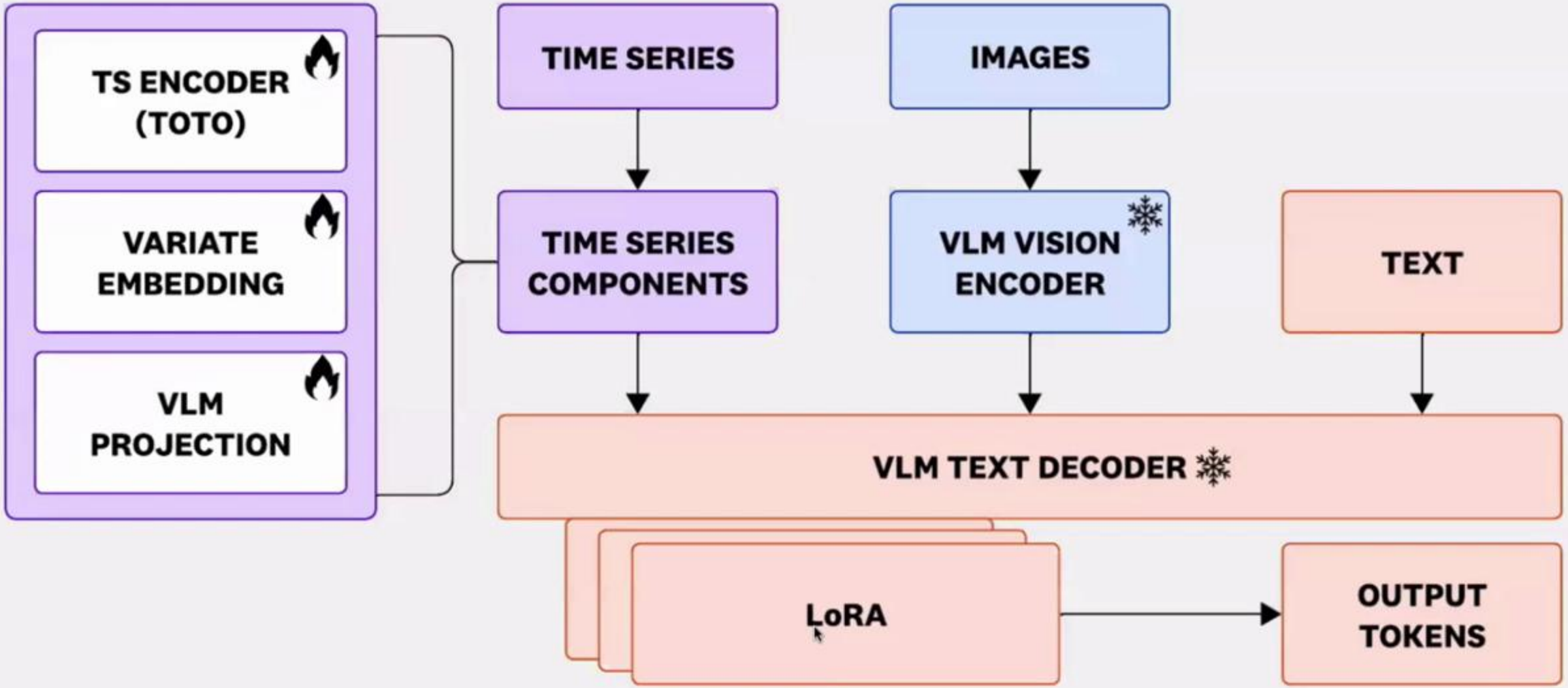
DATADOG


Overall Accuracy
Overall F1
Model-Expert Oracle
Domain Expert
Non-domain Expert
Random Choice
Toto-1.0-QA-Experimental
GPT-5
GPT-5.4
Gemini 3 Pro
Qwen3-VL-32B
GPT-5 (Text)
Claude Opus 4.6
Qwen3-VL-32B (ZS)
Toto-1.0-Qwen3 32B
Qwen3-32B
GPT-4.1
GPT-4o
Claude Sonnet 4.5
Qwen3-VL-8B (ZS)
ChatTS-8B
OpenTSLM-SP 1B
87.2
72.7
69.7
24.5
63.9
62.7
61.3
58.1
56.9
56.4
54.8
52.8
48.8
47.9
47.9
47.2
47.2
45.3
31.1
0.8
82.8
64.6
60.7
22.5
48.9
51.9
51.4
49.6
46.6
43.8
46.7
45.1
33.9
36.1
44.0
42.4
37.9
34.7
22.1
1.2
0
25
50
75
100
Accuracy (%)
F1 (%)
Model Type
LLM
POST-TRAINED TSFM
VLM

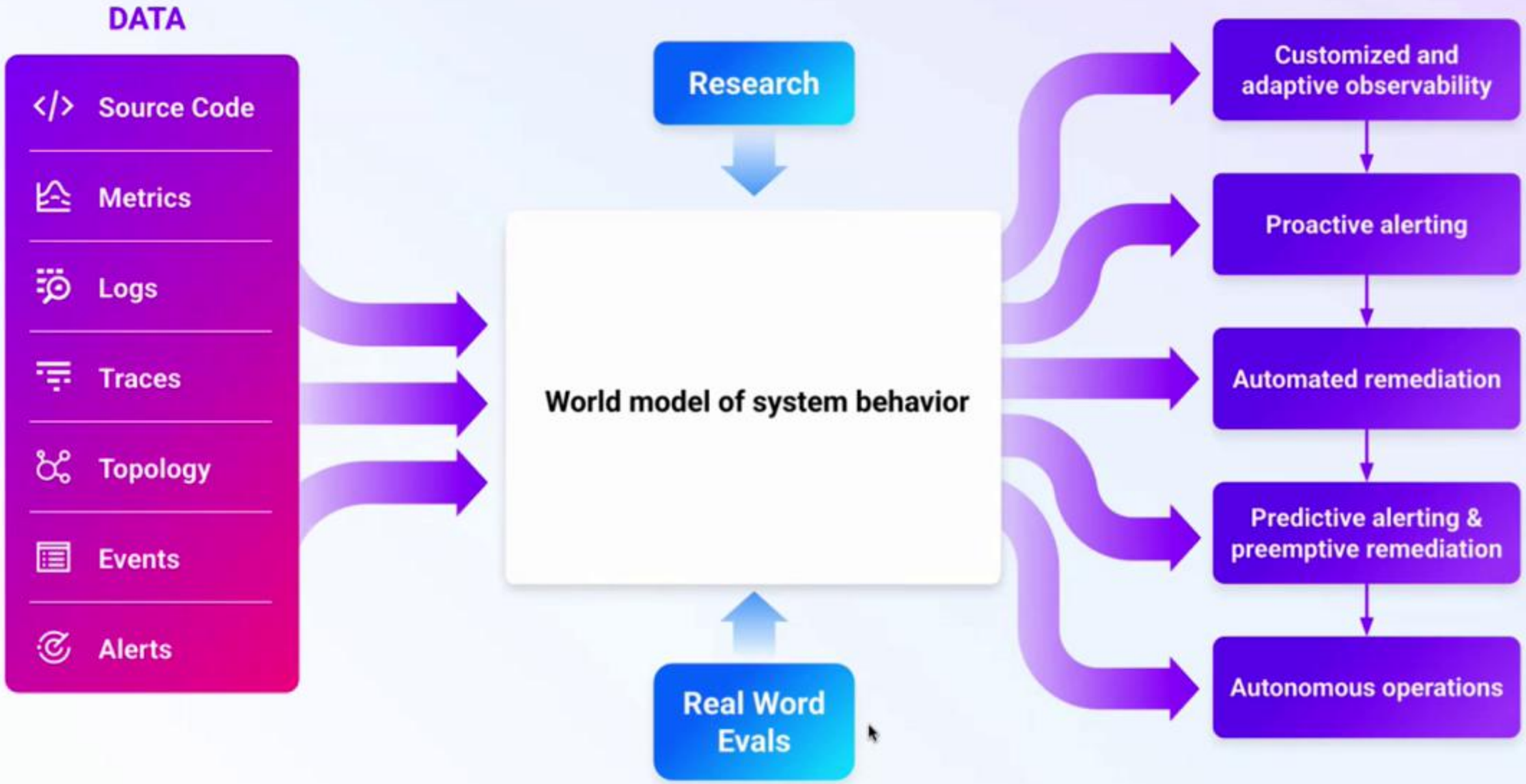
Toto-1.0-QA-Experimental architecture



Toto-1.0-QA-Experimental architecture



Towards autonomous operations




BERT Moment For
Time Series


BERT Moment For
Time Series
GPT-2 Moment For
Time Series

BERT Moment For Time Series

GPT-2 Moment For Time Series

Multimodal World Models for o11y

Datadog AI Research

We are hiring!

Sesh
(VP)

David
(EM II)

Afshin
(Research
Scientist)

Sam
(Research
Scientist)

Bogna
(Engineer)

Arun
(Engineer)

Rocco
(Recruiter)

ICLR

International Conference On
Learning Representations