WELCOME TO

WEBINAR

ODSC UPCOMING EVENTS

AI Builders Summit

Virtual

January 15th - February 8th, 2025

ODSC meetup

Washington DC

February 12th, 2025

Crunching Your Data In-Place With Open-Source LLMs

Webinar

February 19th, 2025

Sean M. Tracey
Head of Developer Relations
Expanso

ODSC meetup

Boston

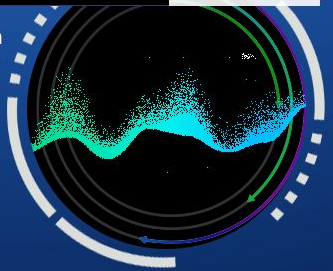
February 27th, 2025

6-Week AI Bootcamp

Virtual

January 6th - February 13th, 2025

Sheamus McGovern
Founder and Engineer
ODSC



ODSC East 2025

Hybrid Event

May 13th - 15th, 2025

Boston

Foundation Models for Time-Series

Stefan Webb, Developer Advocate @ Zilliz

zilliz

Speaker

Stefan Webb

Developer Advocate

milvus zilliz

stefan.webb@zilliz.com

stefan-webb

Why Time-Series?

- Time-series reasoning
  - Social understanding
  - Financial markets
  - Manufacturing processes
  - Weather
  - Etc.

zilliz

CONTENTS

 |

zilliz

1. Foundation Models for Time-Series?

A Foundation Model for Time-Series

- Goal: Zero-shot forecasting of time-series
- Analogous to GPT
  - Decoder-only
  - Scale up:
    - Data
    - Model size
    - Compute for training

But is it possible...?

- “Can large pretrained models trained on massive amounts of time-series data learn temporal patterns that can be useful for time-series forecasting on previously unseen datasets?”
- No vocabulary, grammar, etc.
- Obtaining big data?
- Commonality across time-series?
- In-context prompting emergent behavior?

zilliz

Yes!

A DECODER-ONLY FOUNDATION MODEL FOR TIME-SERIES FORECASTING

A PREPRINT

Abhimanyu Das **Weihao kong** **Rajat Sen** **Yichen Zhou**

Google Research
{abhidas, weihaokong, senrajat, yichenzhou}@google.com

April 19, 2024

ABSTRACT

Motivated by recent advances in large language models for Natural Language Processing (NLP), we design a time-series foundation model for forecasting whose out-of-the-box zero-shot performance on a variety of public datasets comes close to the accuracy of state-of-the-art supervised forecasting models for each individual dataset. Our model is based on pretraining a decoder style attention model with input patching, using a large time-series corpus comprising both real-world and synthetic datasets. Experiments on a diverse set of previously unseen forecasting datasets suggests that the model can yield accurate zero-shot forecasts across different domains, forecasting horizons and temporal granularities.

zilliz

Important: Not a Fine-Tuned LLM!

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

Abstract

Large language models (LLMs) are being applied to time series forecasting. But are language models actually useful for time series? In a series of ablation studies on three recent and popular LLM-based time series forecasting methods, we find that removing the LLM component or replacing it with a basic attention layer does not degrade forecasting performance—in most cases, the results even improve! We also find that despite their significant computational cost, pretrained LLMs do no better than models trained from scratch, do not represent the sequential dependencies in time series, and do not assist in few-shot settings. Additionally, we explore time series encoders and find that patching and attention structures perform similarly to LLM-based forecasters.[1]

zilliz

2. TimesFM
(Google Research, 2024)

Model

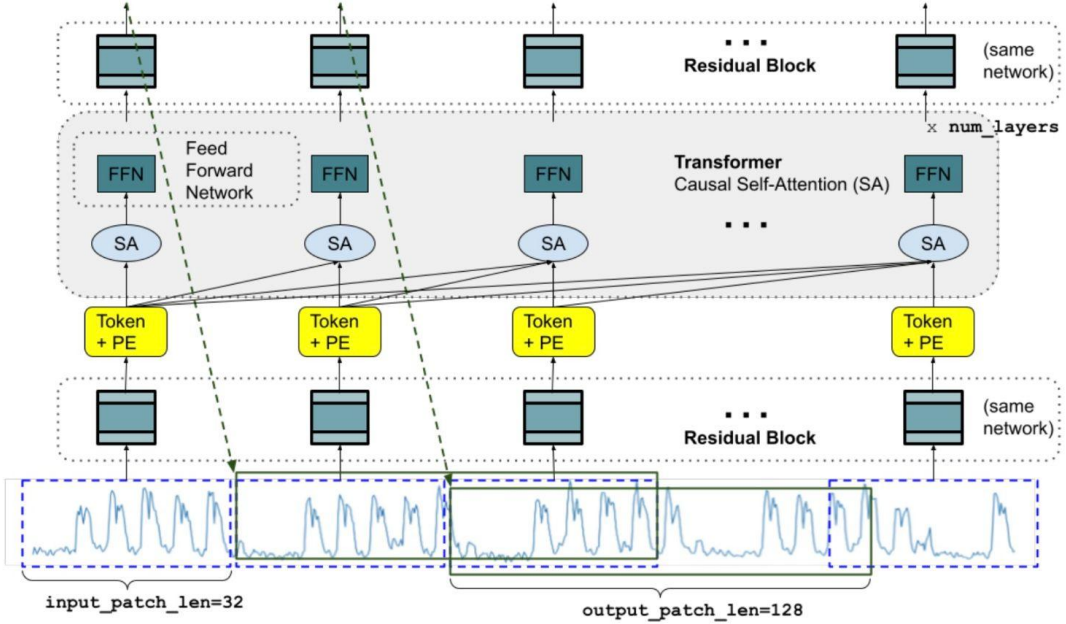


Figure 1: We provide an illustration of the TimesFM model architecture during training, where we show a input time-series of a specific length that can be broken down into input patches. Each patch along is processed into a vector by a residual block (as defined in the model definition) to the model dimension of the transformer layers. The vector is then added to positional encodings and fed into n_l stacked transformer layers. SA refers to self-attention (note that we use multi-head causal attention) and FFN is the fully connected layer in the transformer. The output tokens are then mapped through a residual block to an output of size `output_patch_len`, which is the forecast for the time window following the last input patch seen by the model so far.

zilliz

Data

Google Trends. Google Trends [3] captures search interest over time for millions of queries. We choose around 22k head queries based on their search interest over 15 years from 2007 to 2022. Beyond these head queries the time-series become more than 50% sparse. We download the search interest over time for these queries in hourly, daily, weekly and monthly granularities to form our dataset. The date ranges are Jan. 2018 to Dec. 2019 for hourly and Jan. 2007 to Dec. 2021 for the other granularities. The trends datasets amounts to roughly 0.5B time-points.

Wiki Pageviews. Wiki Pageviews [4] captures the hourly views of all Wikimedia pages. We download all pageview data from Jan. 2012 to Nov. 2023, clean and aggregate the views by page into hourly, daily, weekly and monthly granularities, and filter out pageview time-series with excessive zeros. The final corpus contains roughly 300B time-points.

Synthetic Data. Another major component of our pretraining data is of synthetic origin. We create generators for ARMA [McK84] processes, seasonal patterns (mixture of sines and cosines of different frequencies), trends (linear, exponential with a few change-points) and step functions. A synthetic time-series can be an additive combination of one or more of these processes. We create 3M synthetic time-series each of length 2048 time-points. More details about our synthetic data generation are presented in Appendix A.8.

Other real-world data sources. Along with the wiki and trends data, we also add time-series from several other publicly available datasets to our pretraining corpus. We add all the granularities of the M4 dataset [MSA22], the hourly and 15 minute Electricity and the hourly Traffic datasets (see [ZZP+21]). We also add the 10-minute granularity Weather dataset used for evaluations in [ZZP+21]. M4 has a good mix of granularities with around 100k time-series in total. Traffic and Electricity are large long-term forecasting datasets with > 800 and > 300 time-series each having tens of thousands of time-points. In addition, we add all the 15 min granularity traffic time-series from [WJJ+23].

zilliz

Data

Table 1: Composition of TimesFM pretraining dataset.

Dataset	Granularity	# Time series	# Time points
Synthetic		3,000,000	6,144,000,000
Electricity	Hourly	321	8,443,584
Traffic	Hourly	862	15,122,928
Weather [ZZP+21]	10 Min	42	2,213,232
Favorita Sales	Daily	111,840	139,179,538
LibCity [WJJ+23]	15 Min	6,159	34,253,622
M4 hourly	Hourly	414	353,500
M4 daily	Daily	4,227	9,964,658
M4 monthly	Monthly	48,000	10,382,411
M4 quarterly	Quarterly	24,000	2,214,108
M4 yearly	Yearly	22,739	840,644
Wiki hourly	Hourly	5,608,693	239,110,787,496
Wiki daily	Daily	68,448,204	115,143,501,240
Wiki weekly	Weekly	66,579,850	16,414,251,948
Wiki monthly	Monthly	63,151,306	3,789,760,907
Trends hourly	Hourly	22,435	393,043,680
Trends daily	Daily	22,435	122,921,365
Trends weekly	Weekly	22,435	16,585,438
Trends monthly	Monthly	22,435	3,821,760

zilliz

Training

$$\texttt{TrainLoss} = \frac{1}{N}\sum_{j=1}^{N} \mathrm{MSE}(\hat{\mathbf{y}}_{pj+1:pj+h}, \mathbf{y}_{pj+1:pj+h}).$$

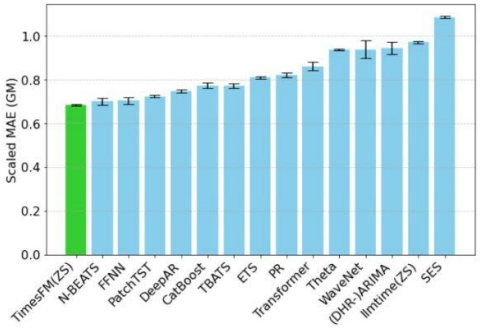
- 200M parameters
  - GPT1 has 120M, GPT2 has 1.5B
- 300B time points
  - GPT1 on BookCorpus, GPT2 on 45M web pages

 |

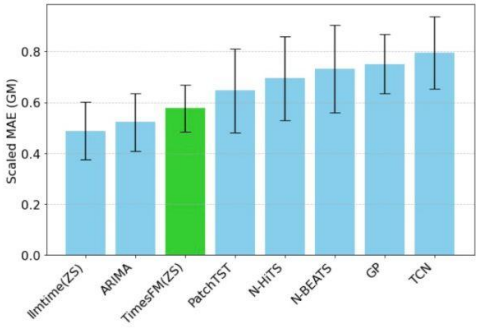
zilliz

3. Discussion

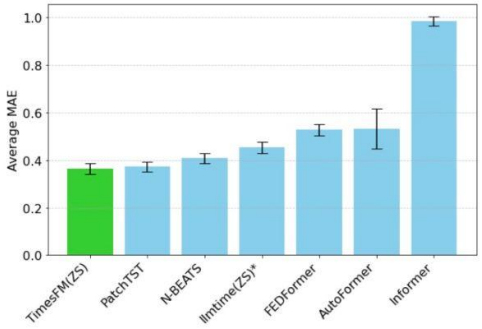
Benchmark Results



(a) Monash Archive [GBW+21]



(b) Darts [HLP+22]

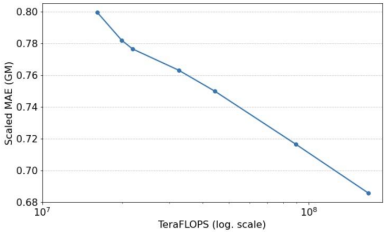


(c) ETT (Horizons 96 and 192) [ZZP+21]

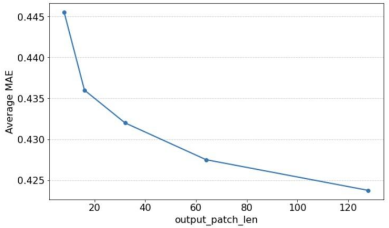
Figure 2: We report average performance in three groups of datasets. In all figures, the lower the metric the better and the error bars represent one standard error. Note that among the baselines only TimesFM and llmtime are zero-shot. In (a) we report results on the Monash datasets. Since the datasets have different scales, we take the Geometric Mean (GM) of the MAE's scaled by the MAE of a naive baseline. We can see that TimesFM is the top model. In (b), we report the similarly scaled MAE on the Darts benchmarks. TimesFM is within significance of the best performing methods which are ARIMA and llmtime in this case. Note that these datasets have one time-series each and therefore statistical methods are competitive with deep learning ones. Finally, in (c) we report the average MAE for 96 and 192 horizon prediction tasks on 4 ETT datasets i.e 8 tasks in total. TimesFM and PatchTST are the best performing models

zilliz

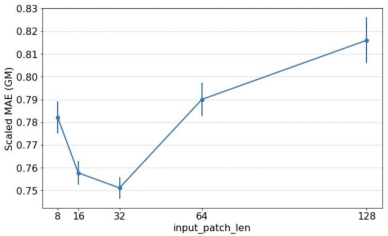
Scaling Laws



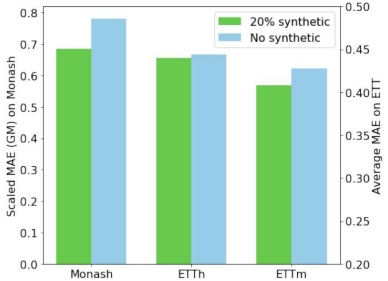
(a) Scaled MAE (GM) on Monash datasets as a function of FLOPS across three model sizes 17M, 70M and 200M. The first 4 points are from 17M and 70M checkpoints while the last 3 are from 200M.



(b) Ablation with respect to output patch length for the task of predicting 512 steps into the future on ETT datasets on the original test set in [ZZP+21]. We report the average across all 4 ETT datasets.



(c) Scaled MAE (GM) for our 70M models on Monash datasets for different input patch lengths. We also plot error bars denoting one standard error.



(d) Average scaled MAE for Monash on the left and average MAE on ETT datasets on the right. We compare the performance of 200M model with and without the synthetic data.

Figure 3: Ablation studies with respect to various design choices.

Conditioning on Exogenous Variables

- Simple method
  - Put regression model on residual
    - Y = TimesFM(X) + f(X, C)
- Exo. vars effecting attention or feedforward networks
- Fine-tuning
  - LoRA weights a function of exo. vars?

 |

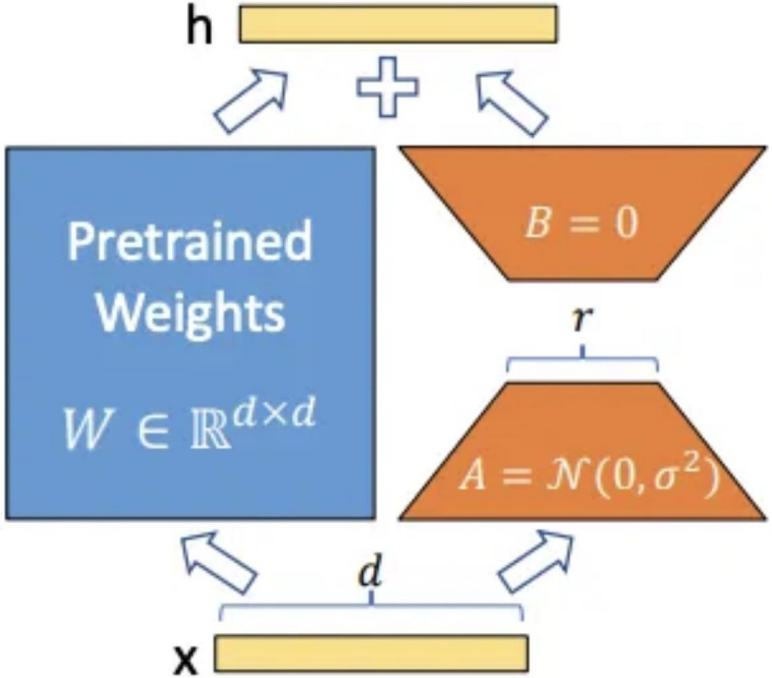
zilliz

Multimodal Foundation Time-Series Models

- Text, images, etc. + time-series ⇒ forecast
  - Include in residual regression model?
  - Include in cross-attention input?
  - Include in LoRA weights hypernet?

zilliz

Fine-Tuning?



 |

zilliz

Challenges / Future Directions

- Data, data, data [hard]
- Exogenous variables and multimodal models, both conditioning on and **generating**
- Structured state-space models and xLSTM
- Embedding models for time-series
    - Decoder-only models can be converted to encoders
- Better tooling
- Connections to Neural Process family?

zilliz

4. RAG / Agents + Time-Series?

Using TimesFM in RAG / Agent Pipeline

- Zero-shot ⇒ Few-shot reasoning, i.e. RAG
- Natural language reasoning and multimodal modeling (multiple modalities in ⇒ time-series out)
- “Agentic workflows” ⇒ Take output of forecast and use as input to generation of text or image
- How does this relate to Milvus?

zilliz

4½. A Brief Interlude on Milvus

Milvus: High-performance, scalable vector database

Milvus is an **Open-Source Vector Database** to **store**, **index**, **manage,** and **use** the massive number of **embedding vectors** generated by deep neural networks and LLMs.

milvus
LF AI &DATA

400+ contributors

32K+ stars

66M+ docker pulls

3K+ forks

zilliz

Milvus Users

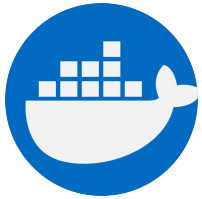
 |

zilliz

Deployment Options

Milvus Lite

- Locally hosted
- Suitable for prototyping and demos



Milvus Standalone

- Single remote/local server
- “Medium” scale
- Simplified setup, maintenance, etc. compared to cluster

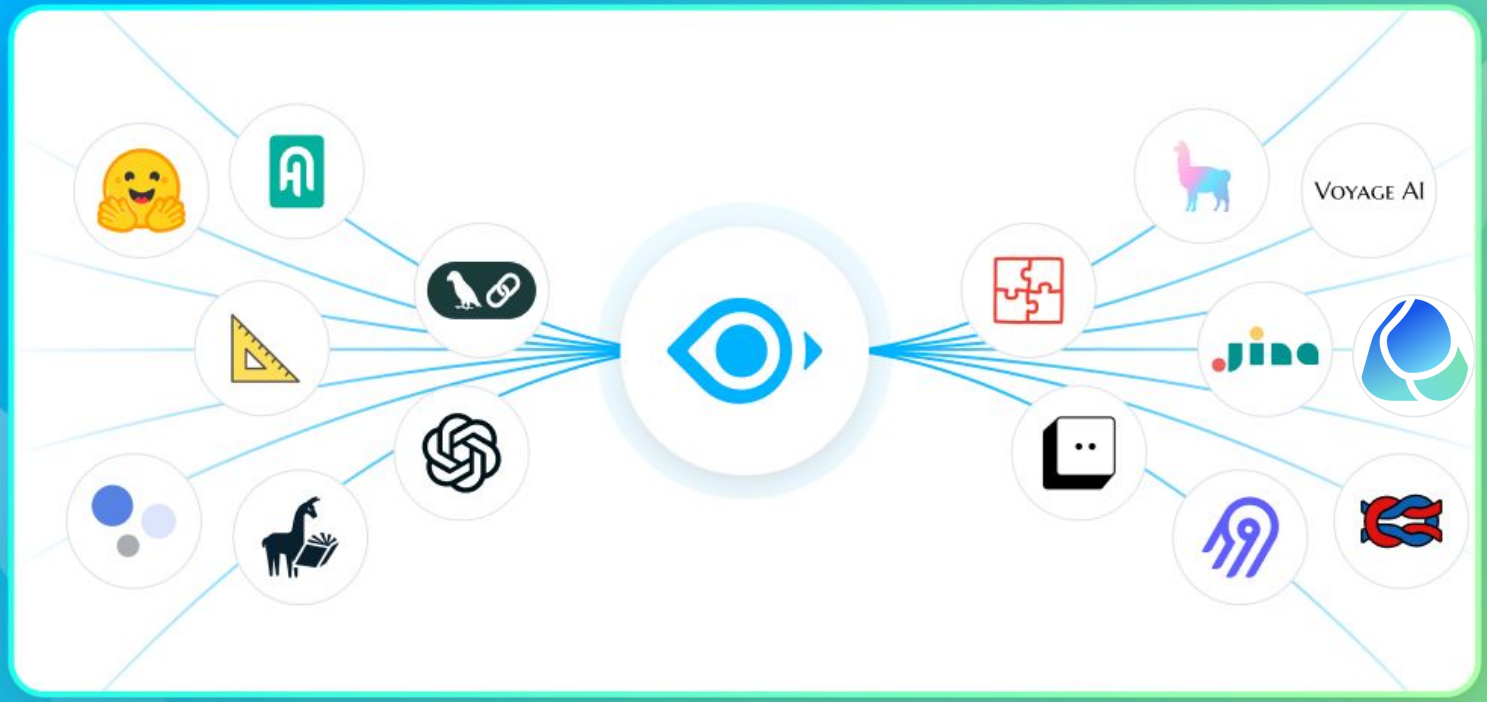
Milvus Cluster

- Distributed system
- Many different types of nodes
- Scales to 100s of billions of vectors

 |

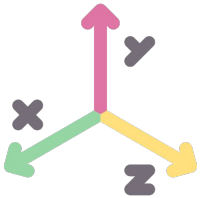
zilliz

Seamless integration with all popular AI toolkits

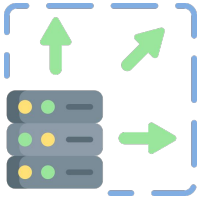


milvus

Why Not Traditional Databases?



Suboptimal Indexing / Search



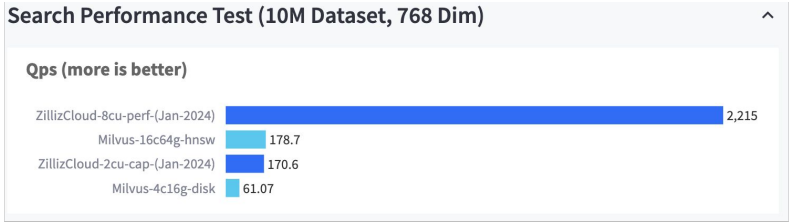
Scaling

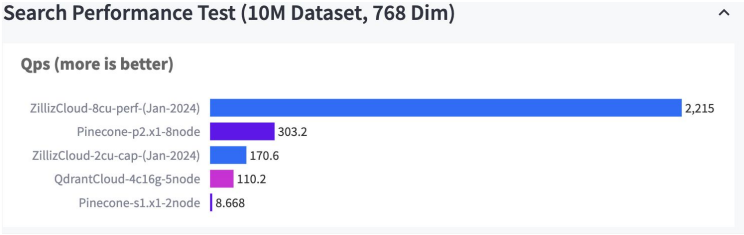
Inadequate Query & Analytics Support

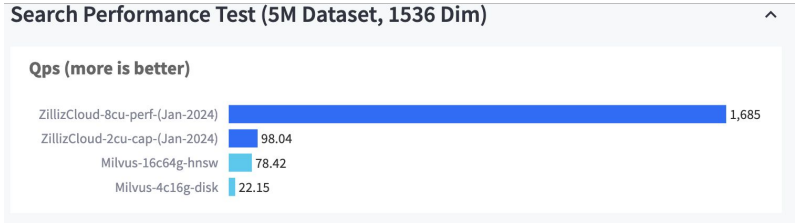
 |

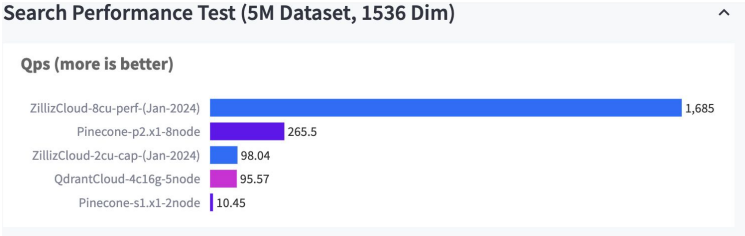
zilliz

Benchmarks









Shows 3-20x faster comparing with open source Milvus

At least 6x faster than other vector databases

https://github.com/zilliztech/VectorDBBench

zilliz

Summary

Summary

- Decoder-only Transformer models show promise for zero-shot time-series forecasting
- It’s early days - stay tuned!
- Many opportunities for Applied Research, but can start building now with Open-Source models
- Better tooling including support for streaming (time-series) data will follow

 |

zilliz

LET'S STAY CONNECTED!

Stefan Webb

Developer Advocate, Zilliz

https://milvus.io/discord

https://github.com/milvus-io/milvus

https://x.com/milvusio

https://www.linkedin.com/company/the-milvus-project

 |

zilliz

Book a free 1:1 session to get help with your production deployment
meetings.hubspot.com/chloe-williams1/milvus-office-hours

Join us at our next meetup!
lu.ma/unstructured-data-meetup

Feb 27, San Francisco, AWS GenAI Loft
Zilliz, Amazon AWS, [TBC]

THANK YOU