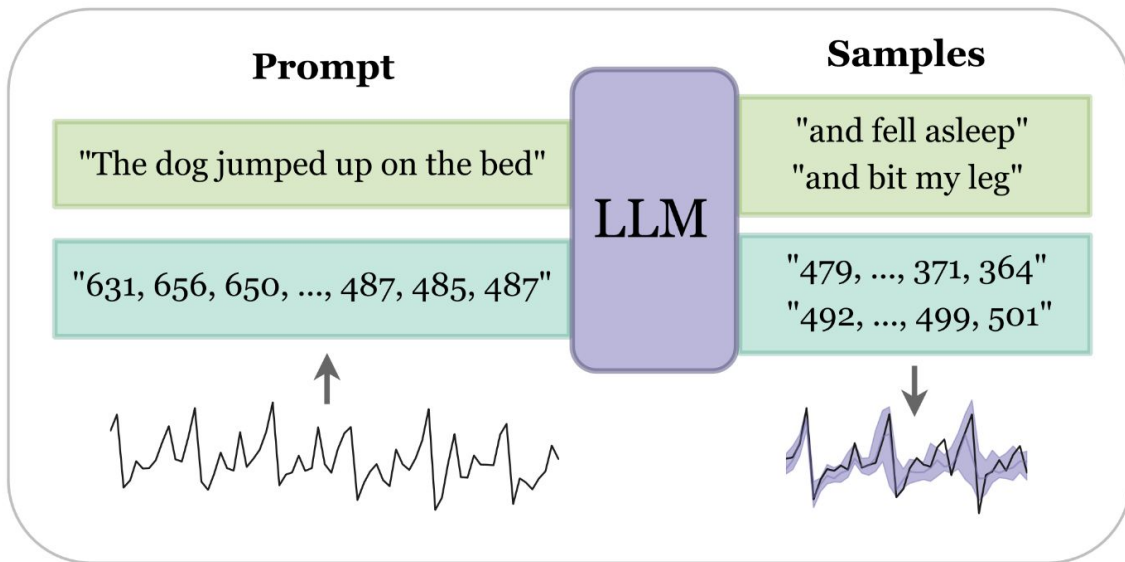
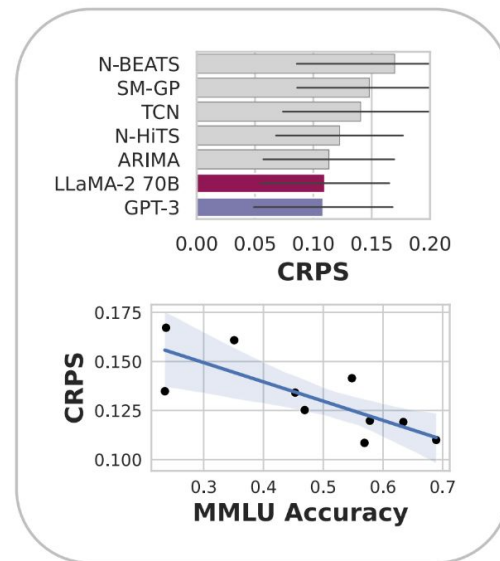
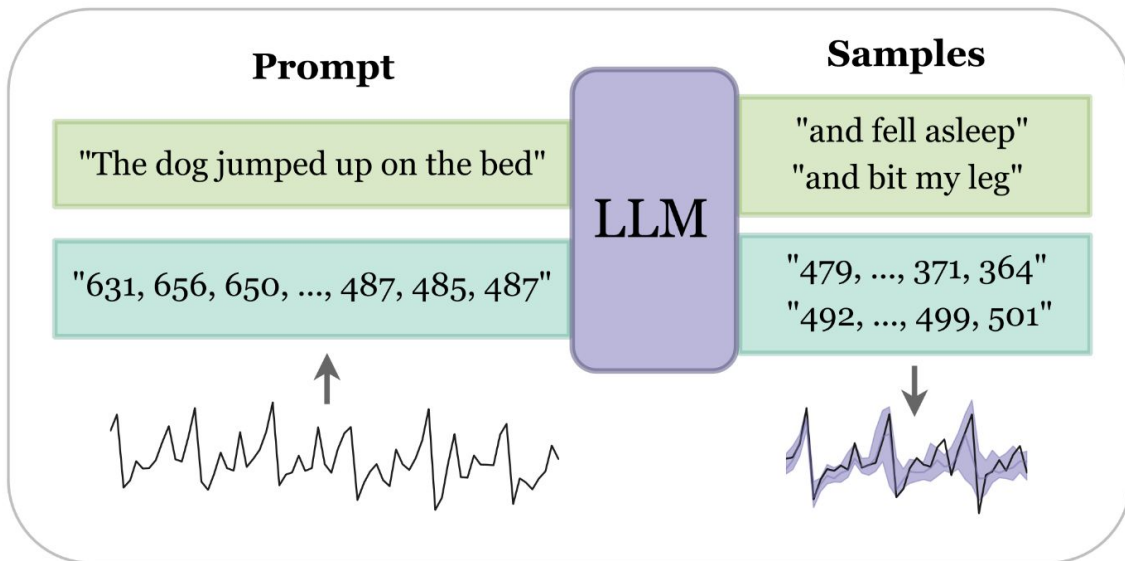


Large Language Models Are Zero-Shot Time Series Forecasters

Nate Gruver*, Marc Finzi*, Shikai Qiu*,
Andrew Gordon Wilson





Outline

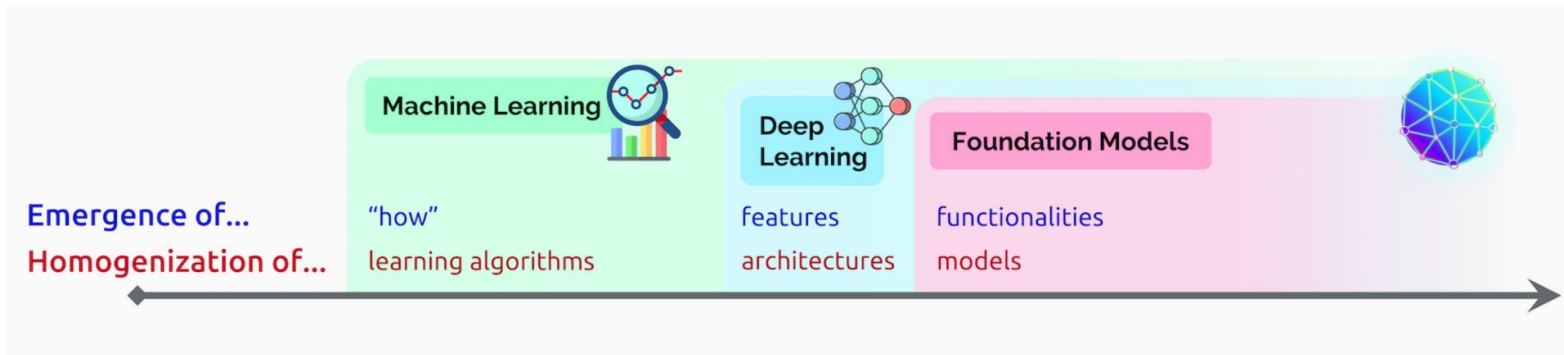
- ❑ Motivating large language models (LLMs) for forecasting
- ❑ Our method (LLMTime)
- ❑ Experimental validation
- ❑ Understand why LLMs work for forecasting
- ❑ Future work

Outline

- ❑ **Motivating large language models (LLMs) for forecasting**
- ❑ Our method (LLMTime)
- ❑ Experimental validation
- ❑ Understand why LLMs work for forecasting
- ❑ Future work

Motivation

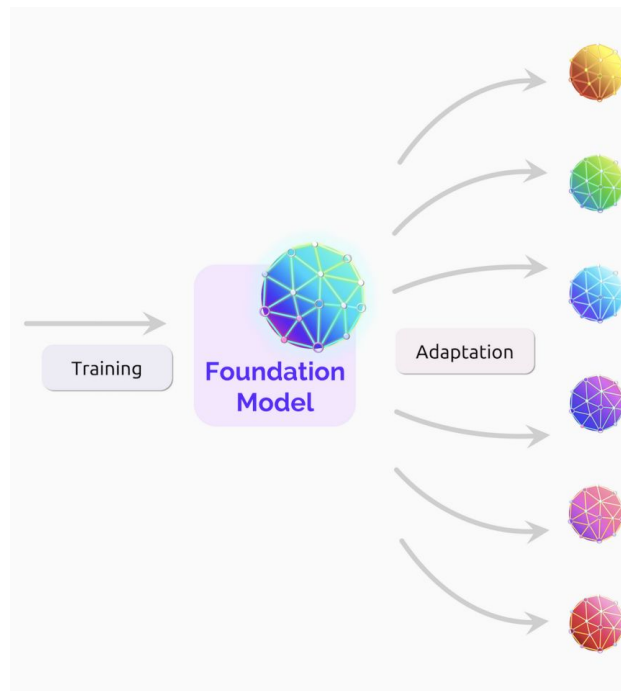
- Foundation models



Motivation

- ❑ Foundation models
 - ❑ ImageNet pretraining
 - ❑ BERT

- ❑ Diffusion models
- ❑ Large language models



Motivation

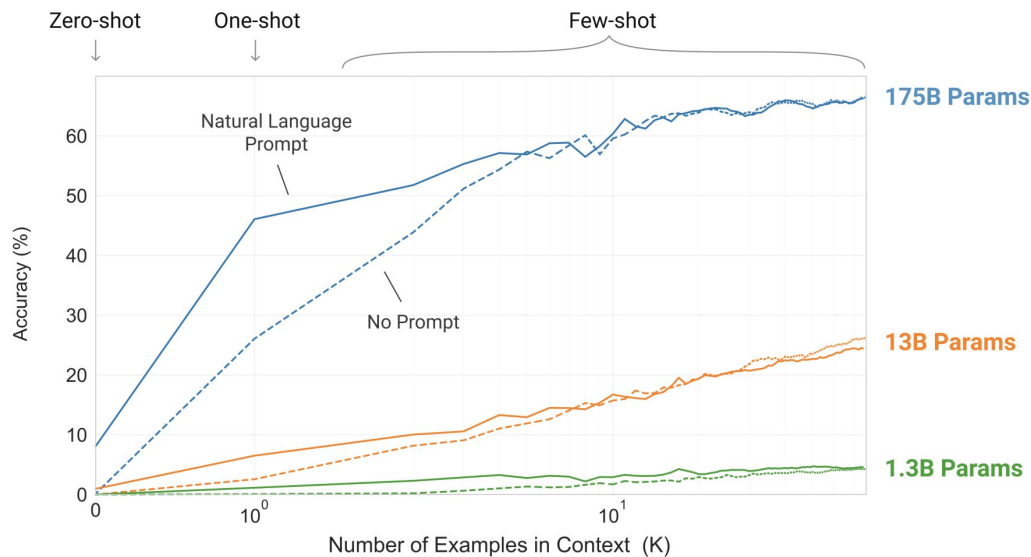
- ❑ Large language models

Context (human-written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Motivation

Large language models



Motivation

- ❑ No foundation models for time series
 - ❑ Models trained from scratch on individual datasets (no transfer)
 - ❑ Models designed uniquely for time series (no generality)

Motivation

- ❑ Challenges of time series
 - ❑ Highly diverse (comes from many different sources)
 - ❑ Highly stochastic (rarely have enough information to predict well)

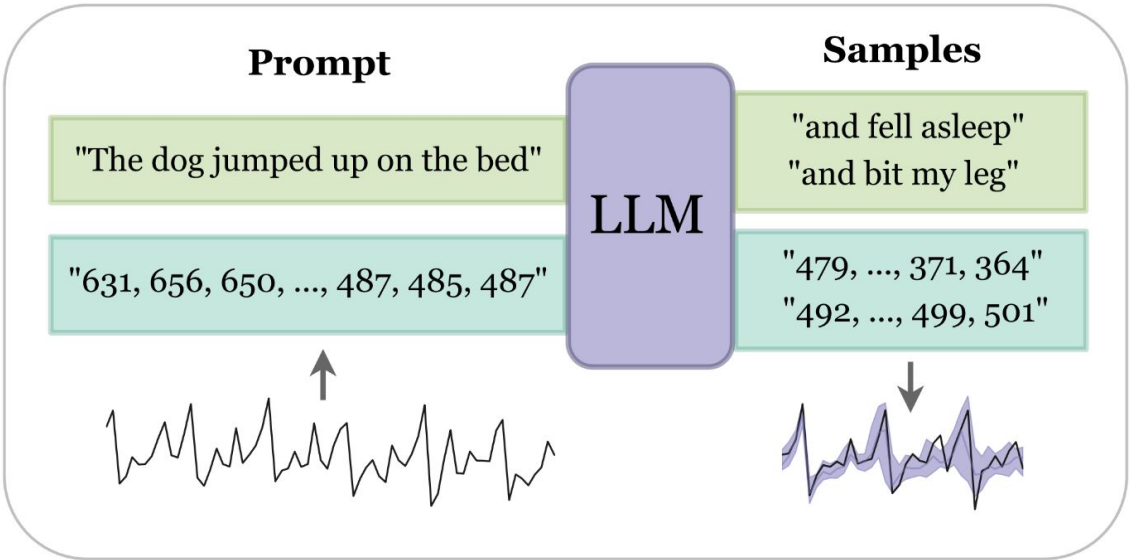
Motivation

- ❑ LLMs might be the right foundation for time series
 - ❑ LLMs are trained on diverse, stochastic sequences (natural language)
 - ❑ LLMs are general, and can fuse time series with text descriptions

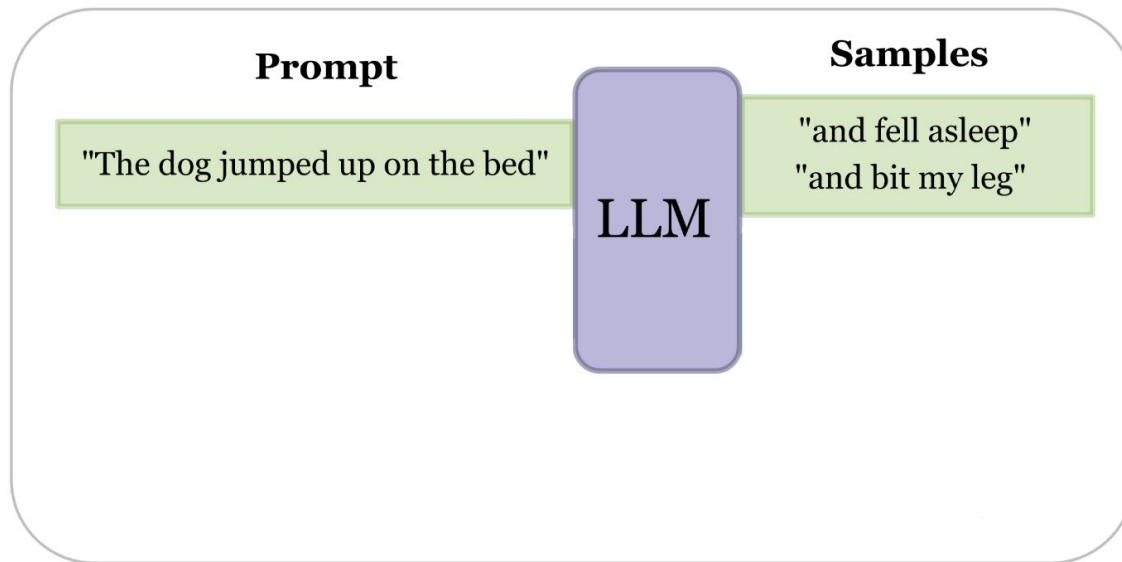
Outline

- ❑ Motivating large language models (LLMs) for forecasting
 - ❑ **Our method (LLMTime)**
- ❑ Experimental validation
- ❑ Understand why LLMs work for forecasting
- ❑ Future work

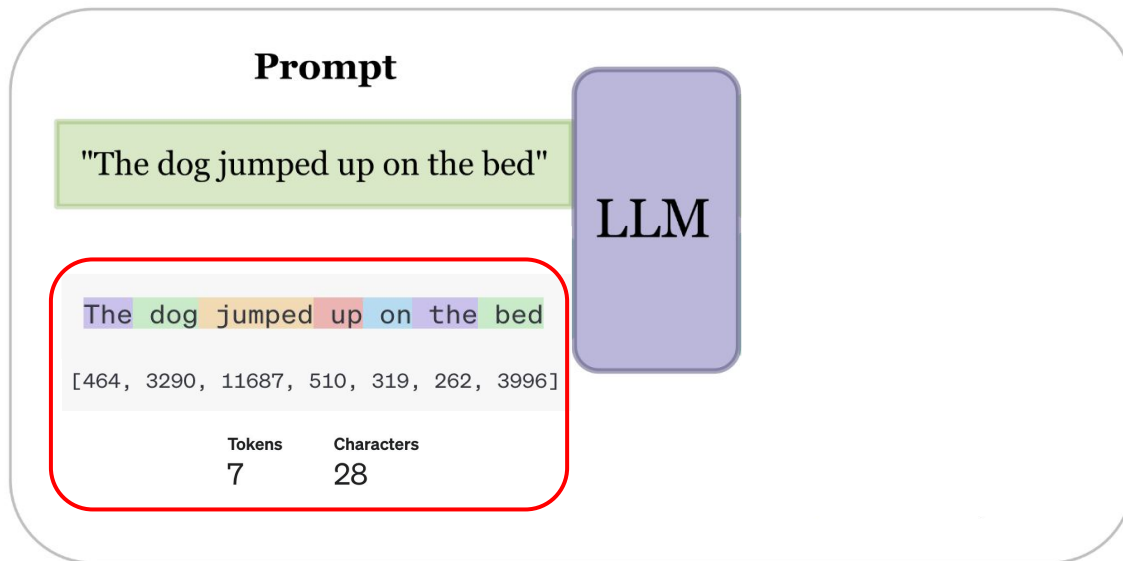
LLMTime



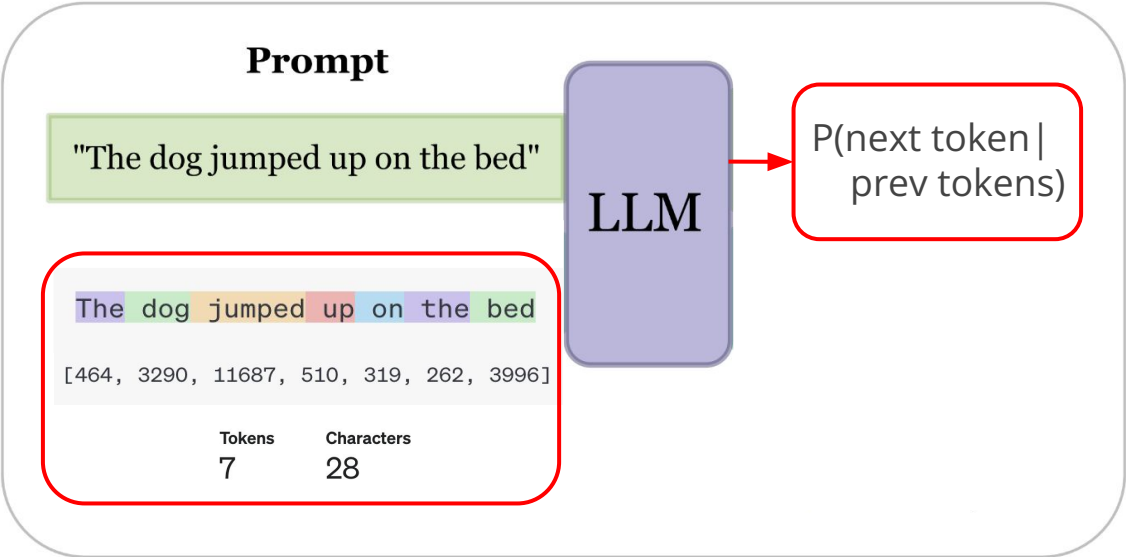
LLMTime - Basics



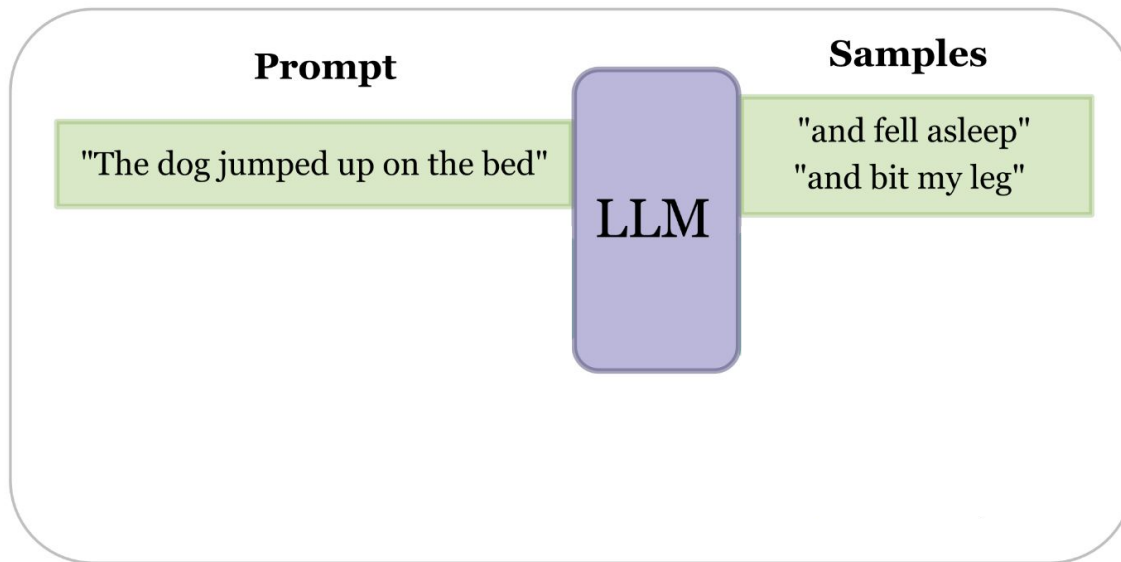
LLMTime - Basics



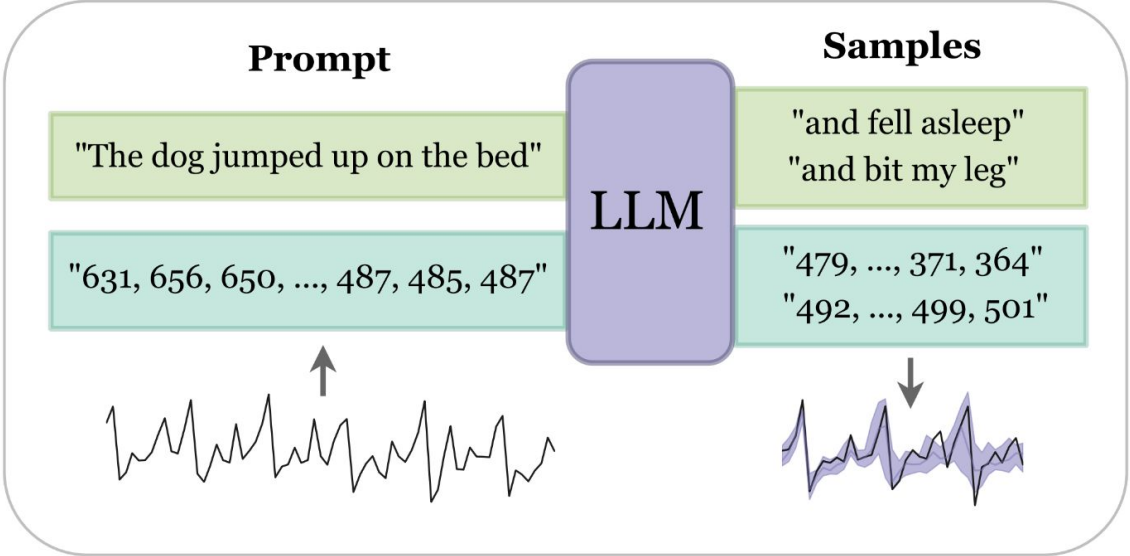
LLMTime - Basics



LLMTime - Basics



LLMTime - Basics



LLMTime - Preprocessing

- ❑ Rescale such that α -percentile is 1
 - ❑ Makes data small, but preserves some variation
- ❑ Offset by β -percent of the range found in the input data
 - ❑ Can make data non-negative

LLMTime - Preprocessing

0.123, 1.23, 12.3, 123.0 → " 1 2 , 1 2 3 , 1 2 3 0 , 1 2 3 0 0 "

LLMTime - Tokenization

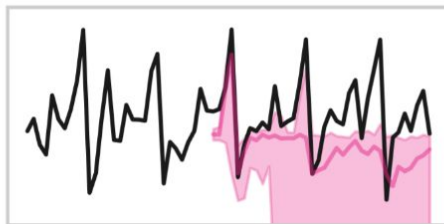
"151,167,...,267"

"151,167,...,267"

LLMTime - Tokenization

"151,167,...,267"

"151,167,...,267"



GPT-3 no spaces

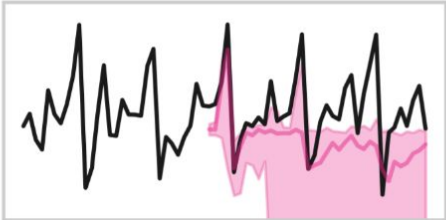
LLMTime - Tokenization

" 1 5 1 , 1 6 7 , ... , 2 6 7 "

"151,167,...,267"

" 1 5 1 , 1 6 7 , ... , 2 6 7 "

"151,167,...,267"

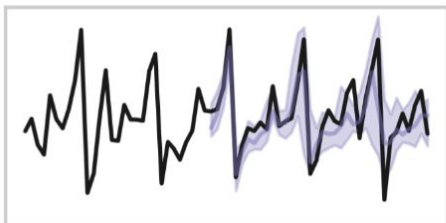


■ GPT-3 no spaces

LLMTime - Tokenization

" 1 5 1 , 1 6 7 , ... , 2 6 7 "

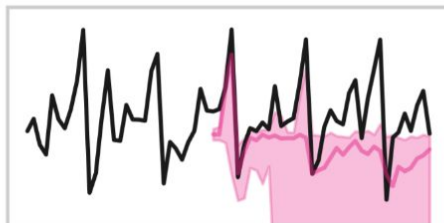
" 1 5 1 , 1 6 7 , ... , 2 6 7 "



■ GPT-3 spaces

"151,167,...,267"

"151,167,...,267"

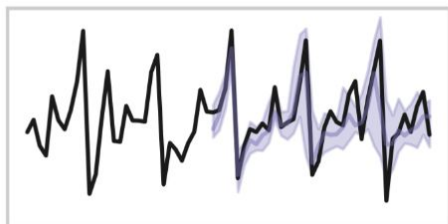


■ GPT-3 no spaces

LLMTime - Tokenization

" 1 5 1 , 1 6 7 , ... , 2 6 7 "

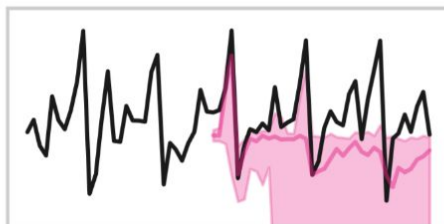
" 1 5 1 , 1 6 7 , ... , 2 6 7 "



■ GPT-3 spaces

"151,167,...,267"

"151,167,...,267"



■ GPT-3 no spaces

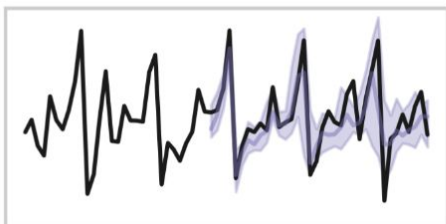
"151,167,...,267"

"151,167,...,267"

LLMTime - Tokenization

" 1 5 1 , 1 6 7 , ... , 2 6 7 "

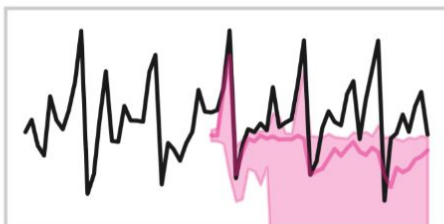
" 1 5 1 , 1 6 7 , ... , 2 6 7 "



■ GPT-3 spaces

"151,167,...,267"

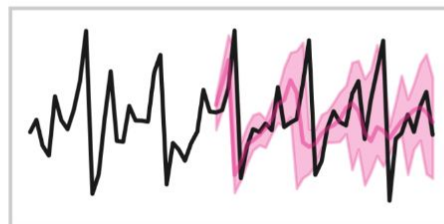
"151,167,...,267"



■ GPT-3 no spaces

"151,167,...,267"

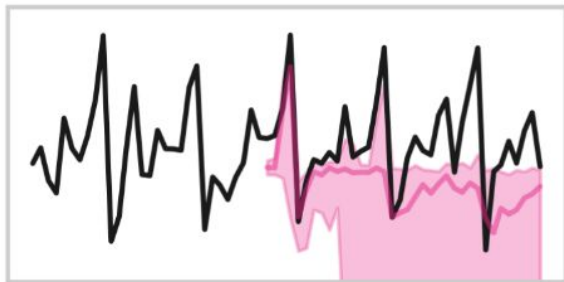
"151,167,...,267"



■ LLaMA no spaces

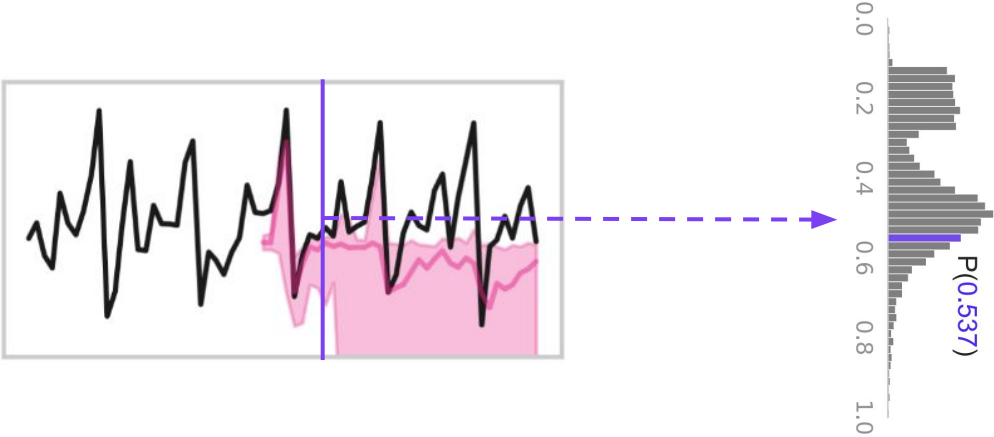
LLMTime - Tokenization

0.123, 1.23, 12.3, 123.0 → " 1 2 , 1 2 3 , 1 2 3 0 , 1 2 3 0 0 "

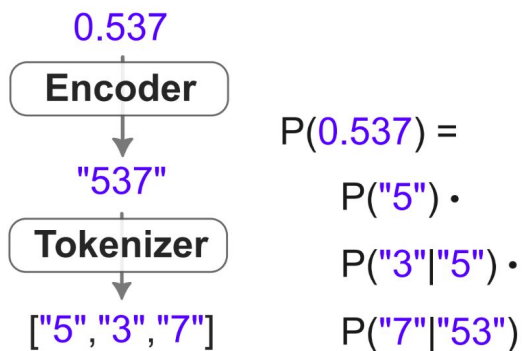


LLMTime - Tokenization

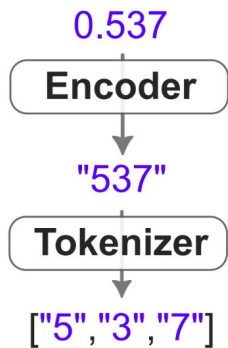
0.123, 1.23, 12.3, 123.0 → " 1 2 , 1 2 3 , 1 2 3 0 , 1 2 3 0 0 "



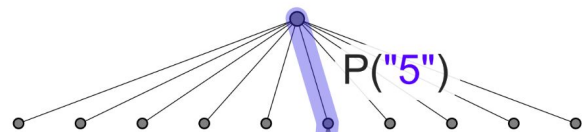
LLMTime - Next Token Prediction



LLMTime - Next Token Prediction

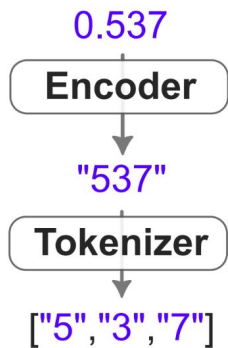


$$P(0.537) = P("5") \cdot P("3|"5") \cdot P("7|"53")$$

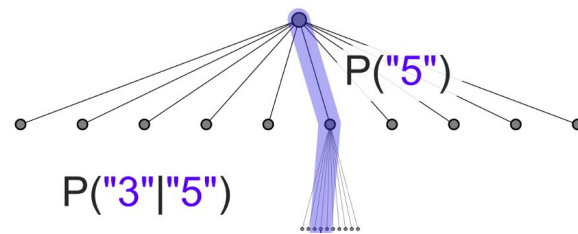


0.0 0.2 0.4 0.6 0.8 1.0

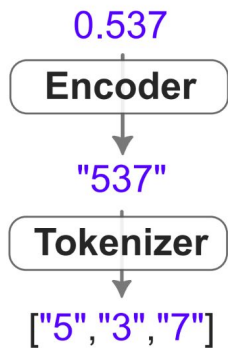
LLMTime - Next Token Prediction



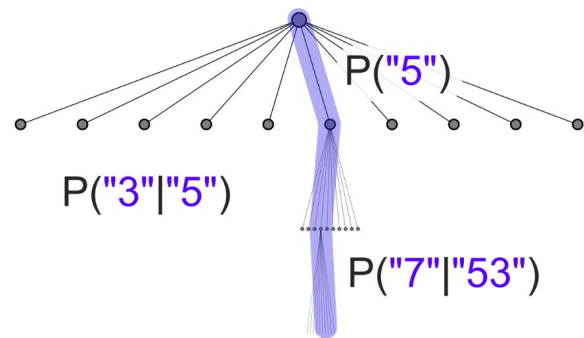
$$P(0.537) = P("5") \cdot P("3|"5") \cdot P("7|"53")$$



LLMTime - Next Token Prediction

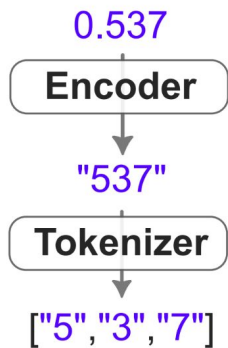


$$P(0.537) = P("5") \cdot P("3|"5") \cdot P("7|"53")$$

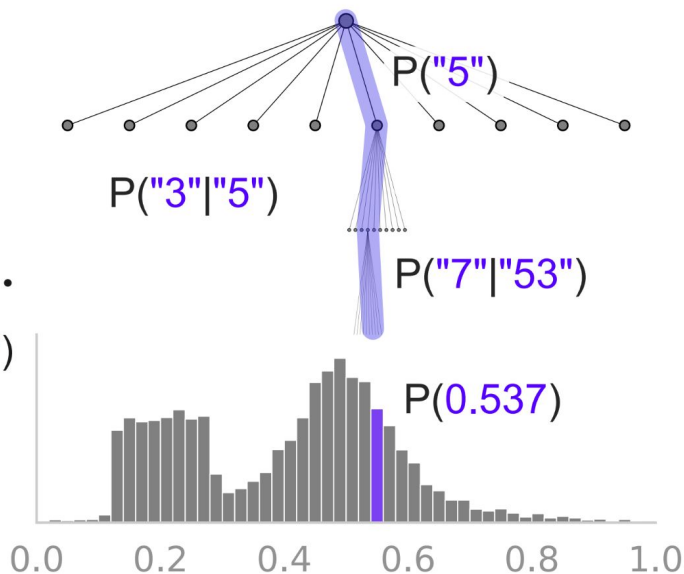


0.0 0.2 0.4 0.6 0.8 1.0

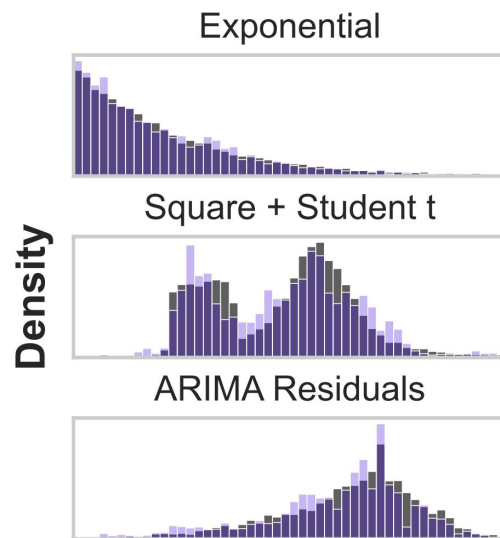
LLMTime - Next Token Prediction



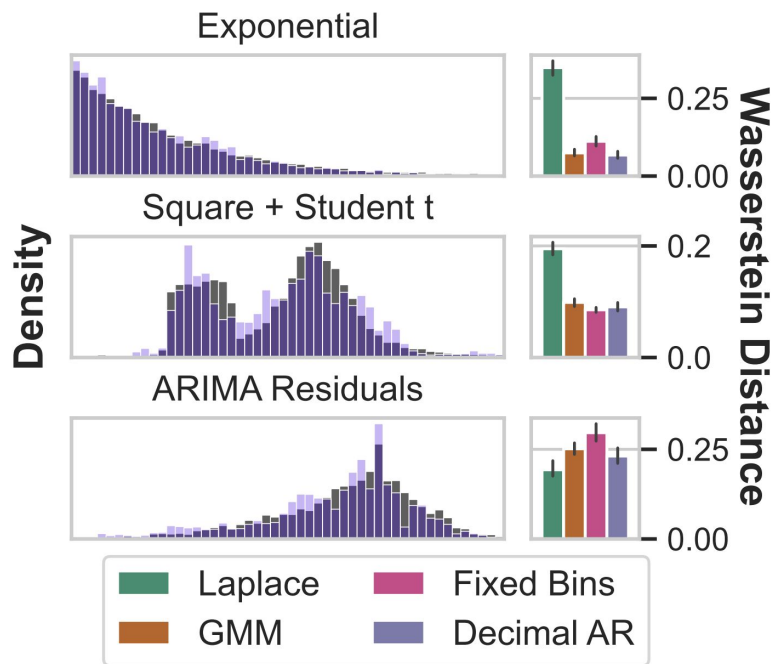
$$P(0.537) = P("5") \cdot P("3"|"5") \cdot P("7"|"53")$$



LLMTime - Next Token Prediction



LLMTime - Next Token Prediction



LLMTime - Summary

- ❑ Simple preprocessing
 - ❑ Rescale and then truncate to fixed precision

LLMTime - Summary

- ❑ Precision is an important hyperparameter
 - ❑ Too small => can't model details
 - ❑ Too large => wasting compute and assigning invalid likelihoods

LLMTime - Summary

- ❑ Simple preprocessing
 - ❑ Rescale and then truncate to fixed precision

- ❑ Careful tokenization
 - ❑ Numbers must map to individual digits

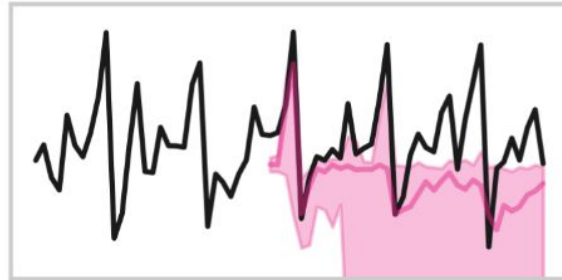
LLMTime - Summary

- ❑ Simple preprocessing
 - ❑ Rescale and then truncate to fixed precision
- ❑ Careful tokenization
 - ❑ Numbers must map to individual digits
- ❑ Continuous probabilistic model
 - ❑ LLMs naturally represent stochastic numbers

Outline

- ❑ Motivating large language models (LLMs) for forecasting
- ❑ Our method (LLMTime)
 - ❑ **Experimental validation**
- ❑ Understand why LLMs work for forecasting
- ❑ Future work

Results - Deterministic Forecasts



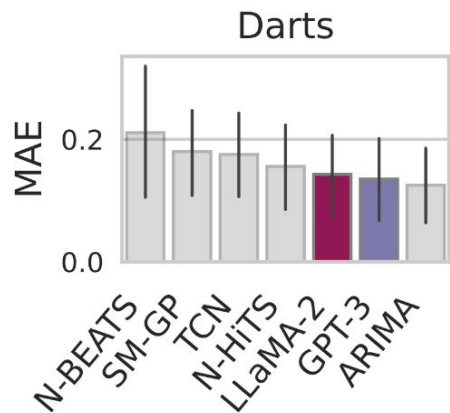
Results - Darts

- ❑ Combines several datasets:
 - ❑ Financial data (sales)
 - ❑ Industrial data (energy usage/production)
 - ❑ Weather data (sunspots)
 - ❑ Healthcare data (vital signs, admissions)

Results - Darts

- ❑ Baseline methods:
 - ❑ ARIMA - traditional method, moving average
 - ❑ TCN - simple convolutional neural network
 - ❑ N-BEATS - neural network designed for time series
 - ❑ N-HiTS - improved version of N-BEATS
 - ❑ SM-GP - Gaussian process with a learned kernel

Results - Deterministic Forecasts



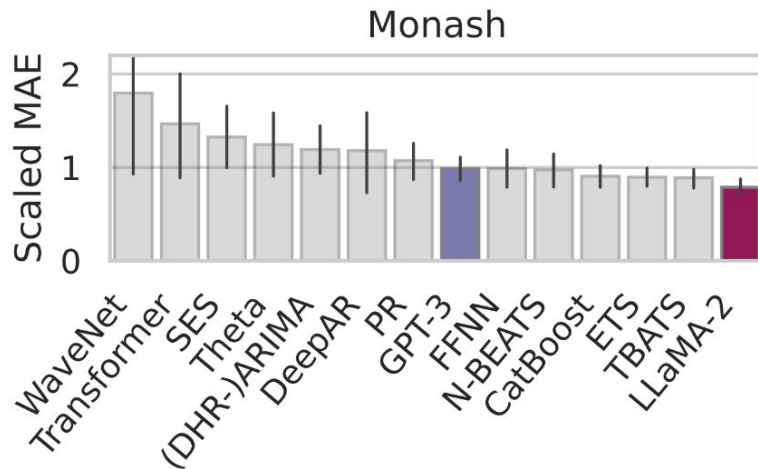
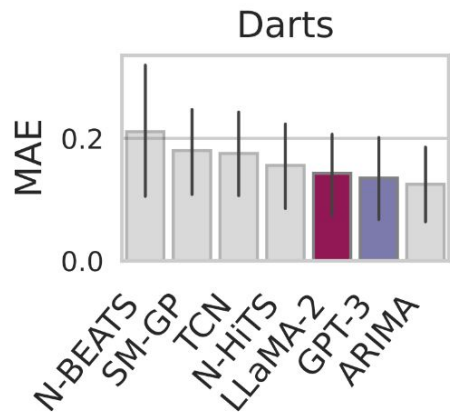
Results - Monash

- ❑ Combines several datasets:
 - ❑ Financial data (tourism)
 - ❑ Industrial data (electricity demand)
 - ❑ Weather data (solar emissions)
 - ❑ Healthcare data (covid deaths)

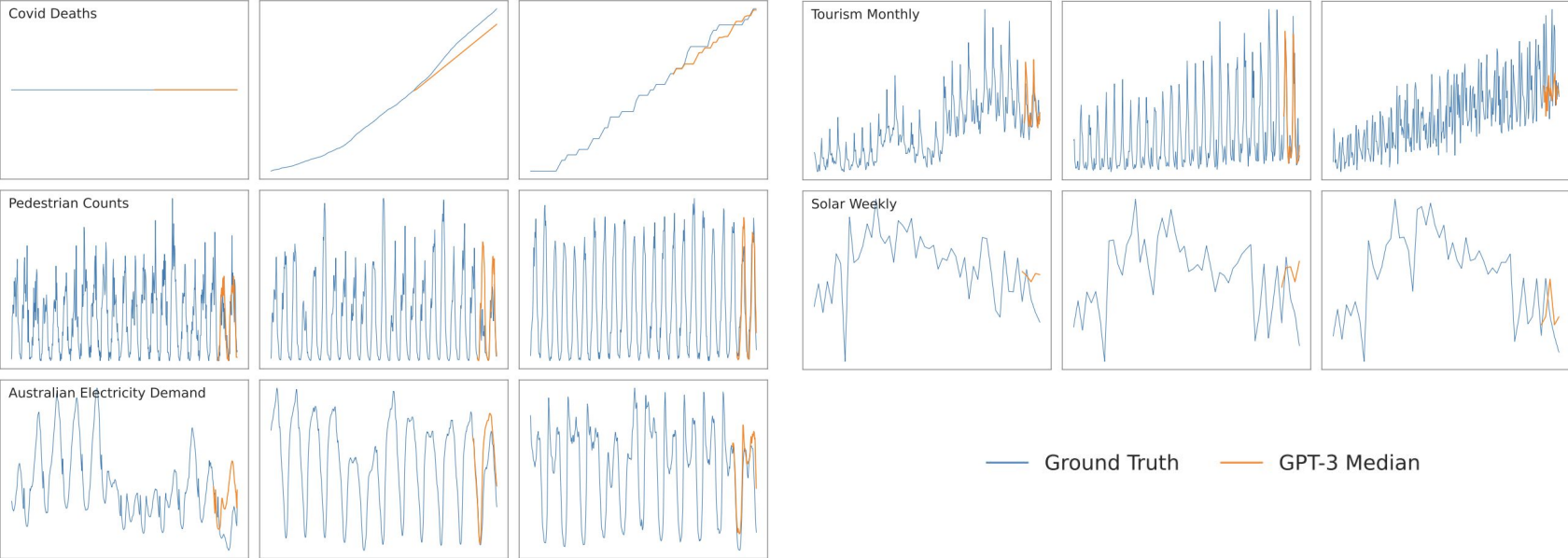
Results - Monash

- ❑ Baseline methods:
 - ❑ Methods used for Darts
 - ❑ CatBoost - widely used data mining method
 - ❑ FFNN - a very simple neural network approach
 - ❑ PR - a variant of linear regression

Results - Deterministic Forecasts



Results - Deterministic Forecasts



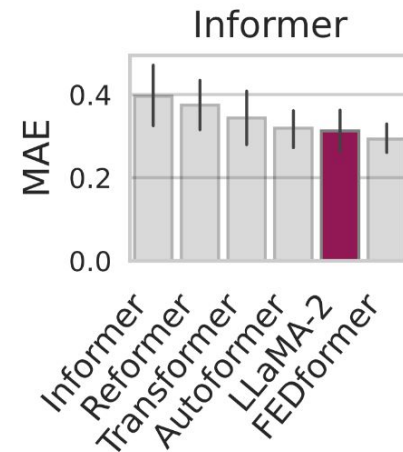
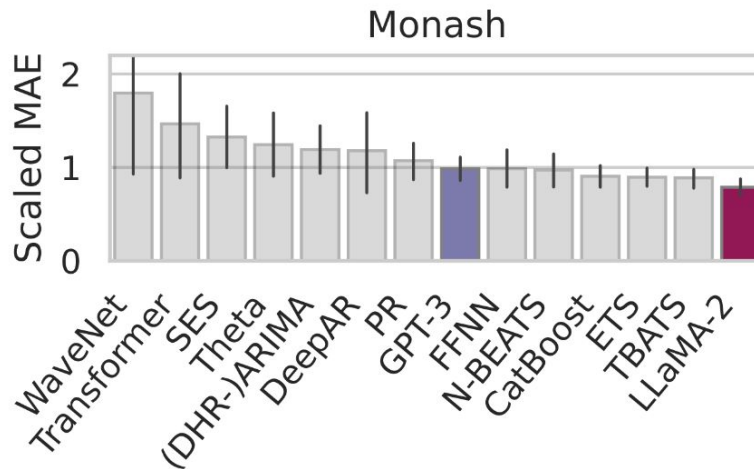
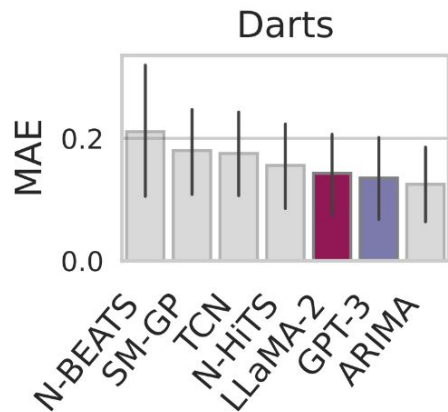
Results - Informer

- ❑ Combines several datasets:
 - ❑ Financial data (exchange rate)
 - ❑ Industrial data (electricity usage, traffic)
 - ❑ Weather data (wind speeds)

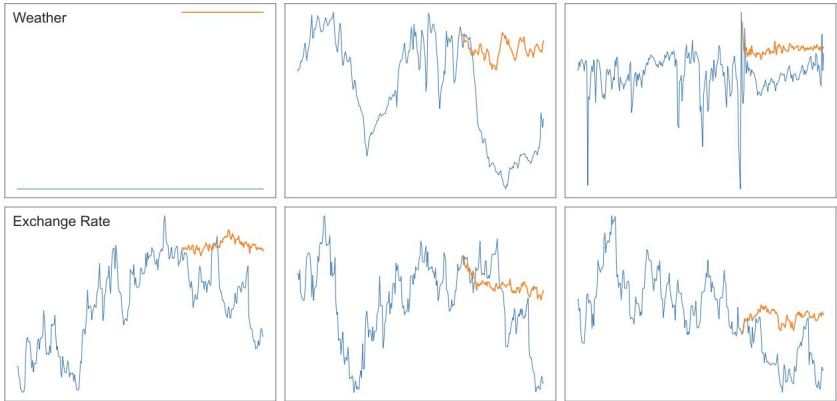
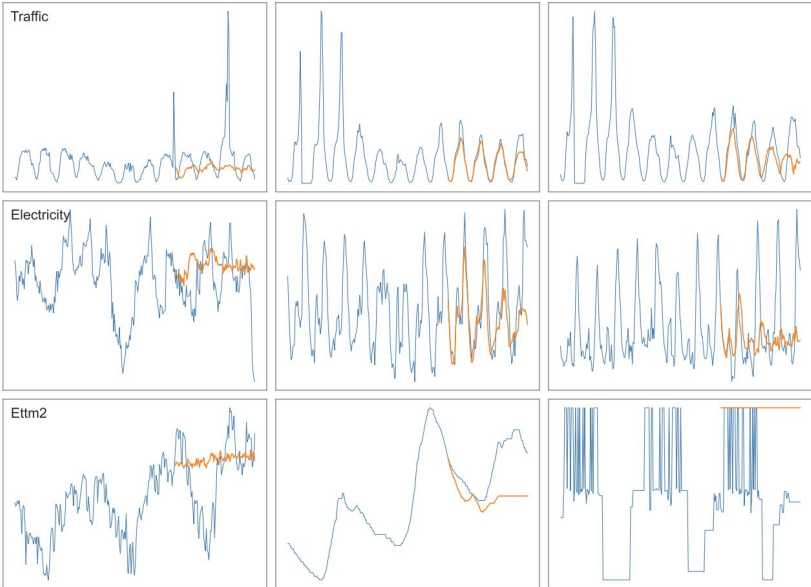
Results - Informer

- ❑ Baseline methods:
 - ❑ Informer - efficient attention mechanism
 - ❑ Reformer - efficient attention mechanism
 - ❑ Transformer - standard architecture
 - ❑ Autoformer - time series inspired architecture

Results - Deterministic Forecasts



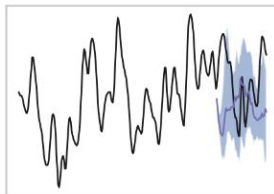
Results - Deterministic Forecasts



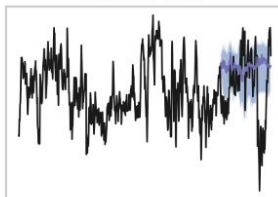
— Ground Truth — GPT-3 Median

Results - Stochastic Forecasts

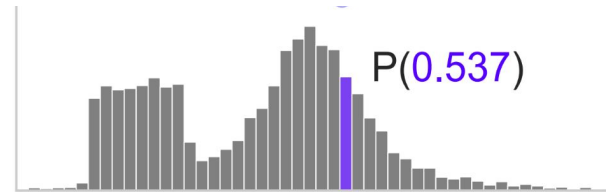
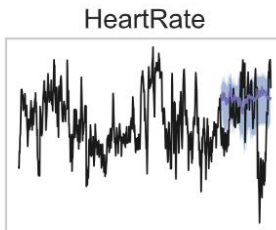
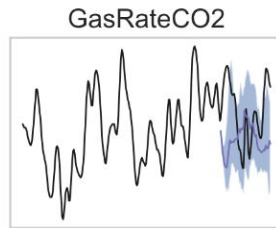
GasRateCO2



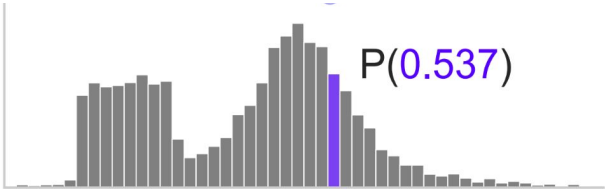
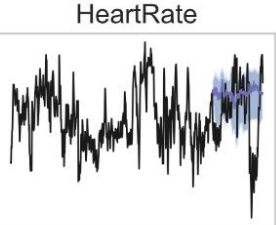
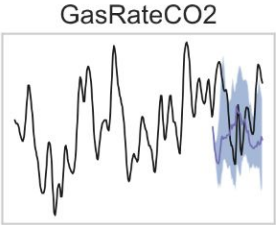
HeartRate



Results - Stochastic Forecasts



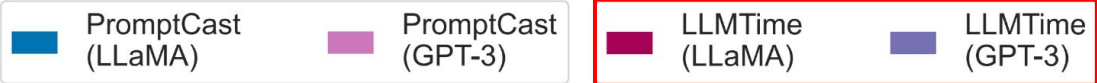
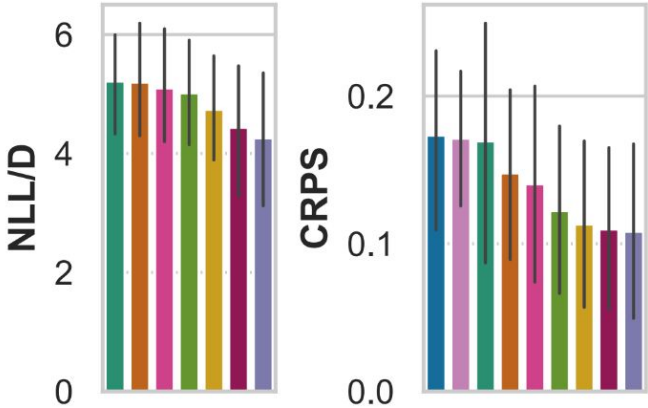
Results - Stochastic Forecasts



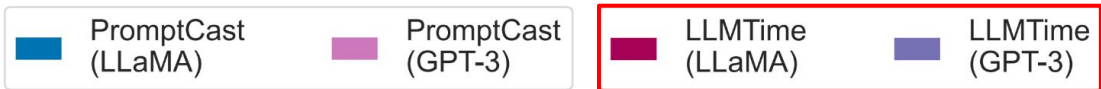
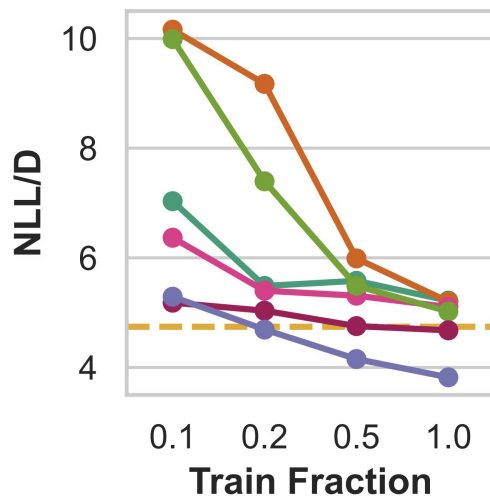
$$\text{CRPS}(\hat{F}, y) = \int_{\mathbb{R}} \left(\hat{F}(z) - \mathbb{I}_{(z-y)>0} \right)^2 dz$$

$\hat{F}(z)$ is the empirical CDF produced by sampling

Results - Stochastic Forecasts



Results - Stochastic Forecasts

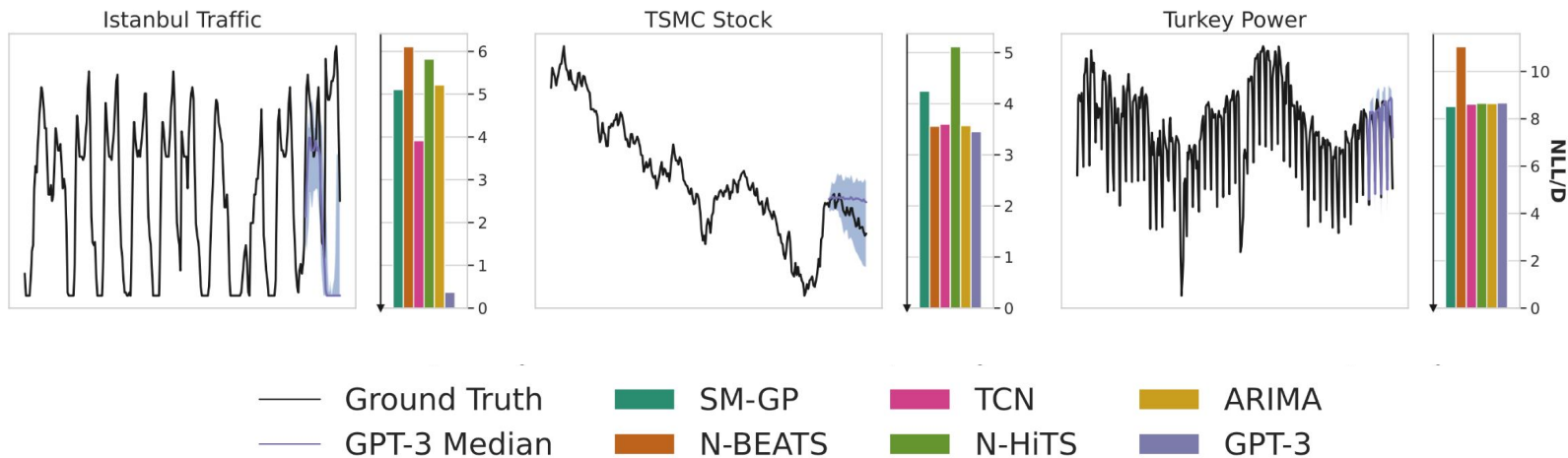


Results - Addressing Memorization

- ❑ How do we know the LLM has seen the test sets before?

Results - Addressing Memorization

□ How do we know the LLM has seen the test sets before?

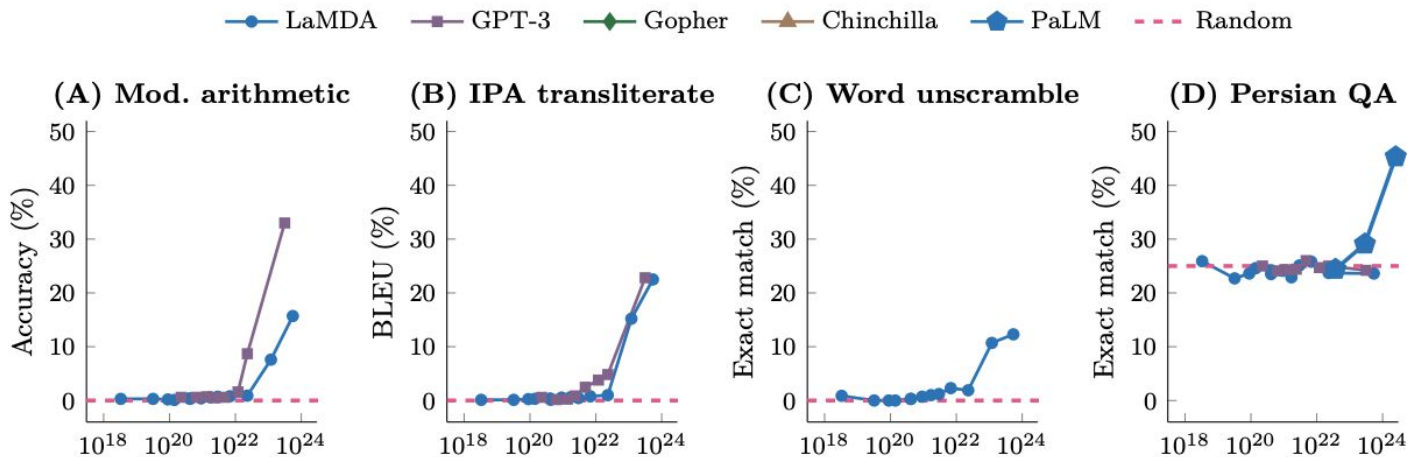


Comparing Models

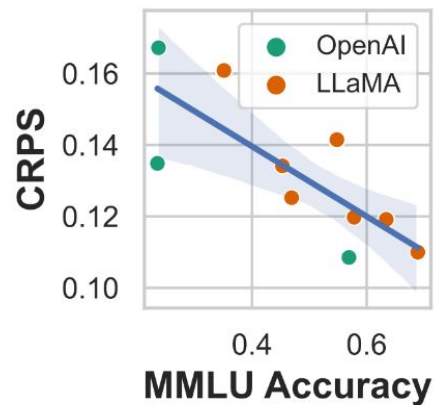
- ❑ Scaling laws

Comparing Models

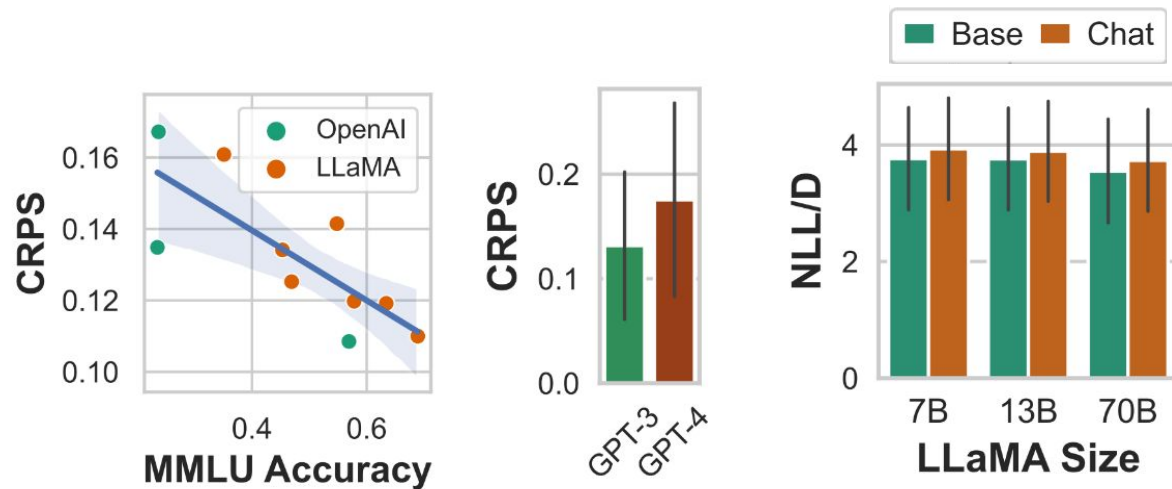
Scaling laws



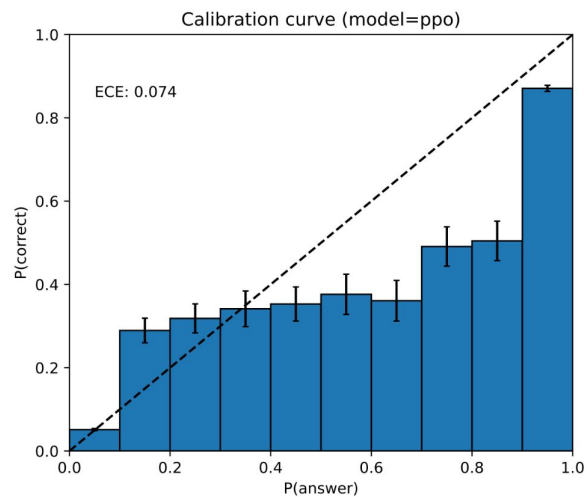
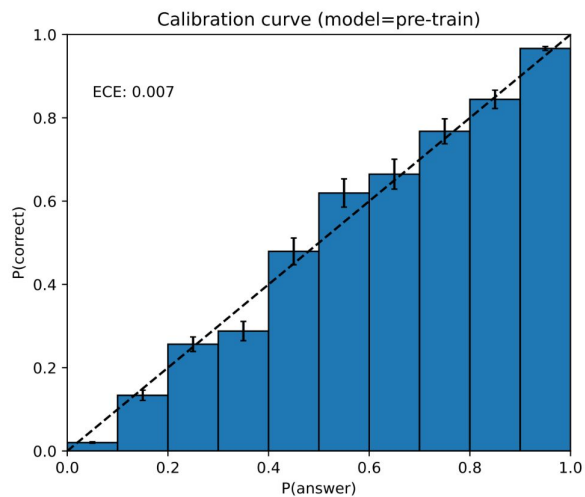
Comparing Models



Comparing Models



Comparing Models



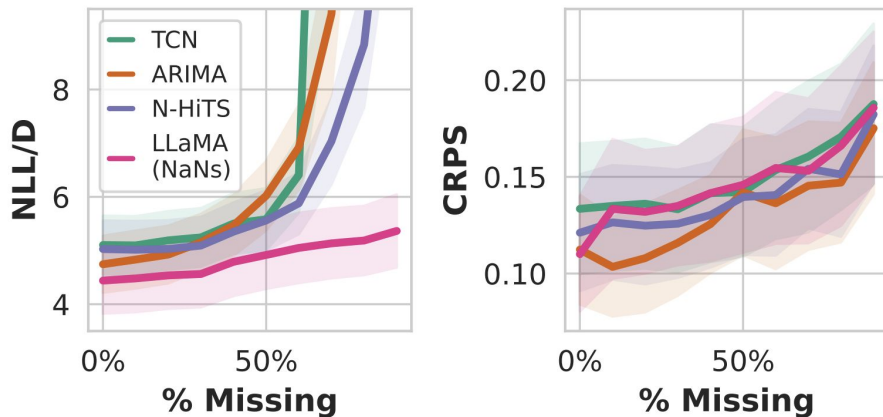
Using Text Conditioning

Using Text Conditioning

[64, , , 49, , 16,] → "64, NaN, NaN, 49, NaN, 16, NaN"

Using Text Conditioning

[64, , , 49, , 16,] → "64, NaN, NaN, 49, NaN, 16, NaN"



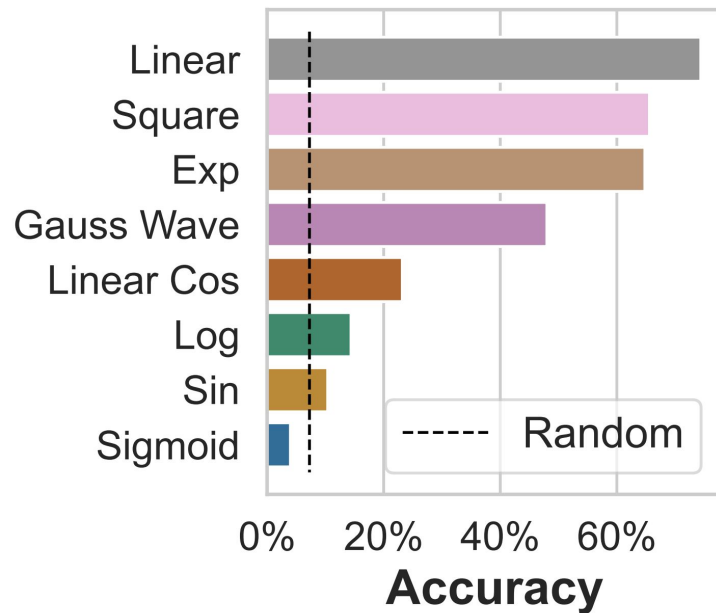
Using Text Conditioning

```
1 First, let's analyze the time series and make some observations:
2
3 1. The time series starts at a negative value, increases linearly, and
   ends at a positive value.
4 2. There is no apparent periodicity or oscillation in the time series.
5 3. There are no sudden jumps or changes in the time series, indicating
   a smooth function.
6
7 Now, let's go through the functions in the mapping and see which one
   matches our observations:
```

Using Text Conditioning

```
15 - 'beat': This function has a periodic oscillation, which doesn't
    match our observations.
16 - 'sigmoid': This function has a sigmoid shape, which doesn't match
    our observations.
17 - 'log': This function increases logarithmically, which doesn't match
    the linear increase in our observations.
18 - 'x_times_sine': This function has a periodic oscillation, which
    doesn't match our observations.
19 - 'square': This function has a parabolic shape, which doesn't match
    our observations.
20
21 Based on our observations and analysis, the function that most likely
    generated the time series is 'linear'.
22
23 Answer: linear
```

Using Text Conditioning

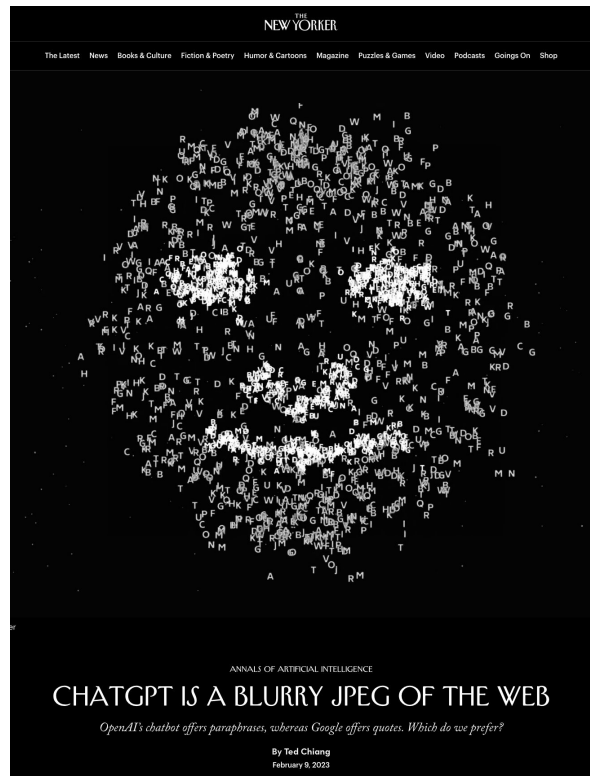


Outline

- ❑ Motivating large language models (LLMs) for forecasting
- ❑ Our method (LLMTime)
- ❑ Experimental validation
 - ❑ **Understand why LLMs work for forecasting**
- ❑ Future work

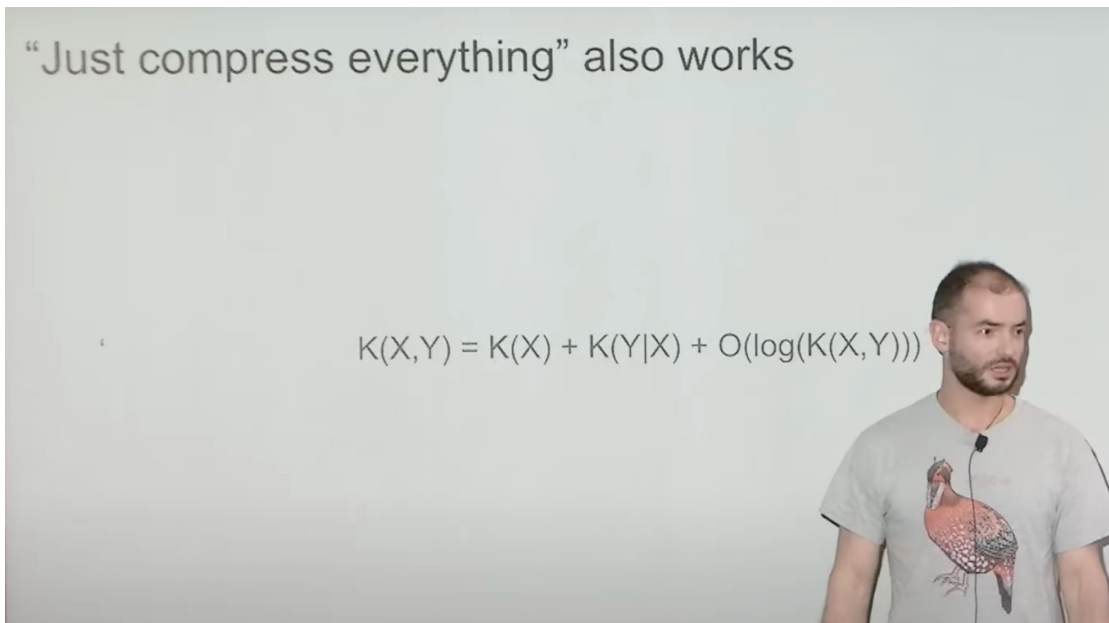
Why Do LLMs Work?

- LLMs are compressors



Why Do LLMs Work?

- LLMs are compressors



Why Do LLMs Work?

- ❑ Compression means finding simplifying explanations
 1. Identify common patterns
 2. Create a rule that explains the pattern

Why Do LLMs Work?

- ❑ Compression means finding simplifying explanations
 1. Identify common patterns
 2. Create a rule that explains the pattern

Example: language models have a bias towards repetition

Why Do LLMs Work?

- LLMs can compress non-text data

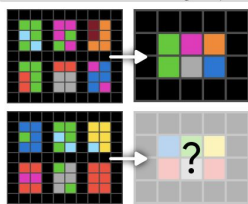
Chunk Size	Compressor	Raw Compression Rate (%)		
		enwik9	ImageNet	LibriSpeech
∞	gzip	32.3	70.7	36.4
	LZMA2	23.0	57.9	29.9
	PNG	42.9	58.5	32.2
	FLAC	89.5	61.9	30.9
2048	gzip	48.1	68.6	38.5
	LZMA2	50.0	62.4	38.2
	PNG	80.6	61.7	37.6
	FLAC	88.9	60.9	30.3
	Transformer 200K	30.9	194.0	146.6
	Transformer 800K	21.7	185.1	131.1
	Transformer 3.2M	17.0	215.8	228.2
	Chinchilla 1B	11.3	62.2	24.9
	Chinchilla 7B	10.2	54.7	23.6
	Chinchilla 70B	8.3	48.0	21.0

Why Do LLMs Work?

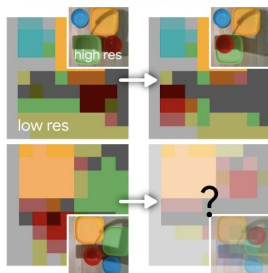
Sequence Transformation

Pattern transformations (symbolic)

Abstraction and Reasoning Corpus



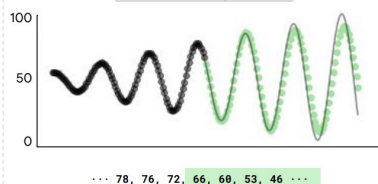
Object Rearrangement



Sequence Completion

Simple function classes (numeric)

Sinusoid Extrapolation



Whiteboard Drawing

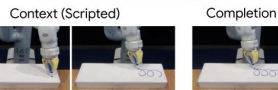
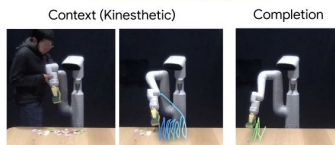


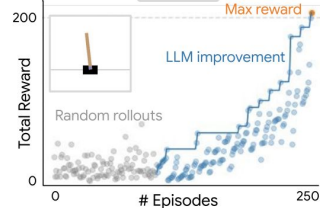
Table Sweeping



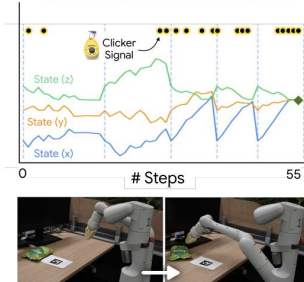
Sequence Improvement

Online policies (numeric & symbolic)

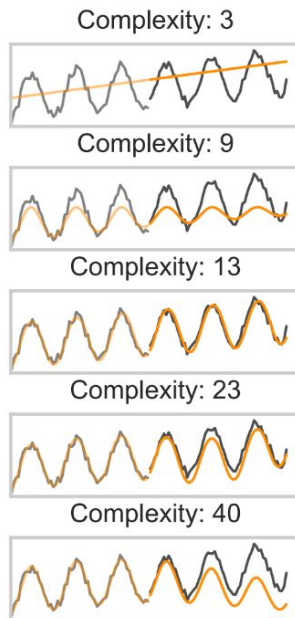
CartPole



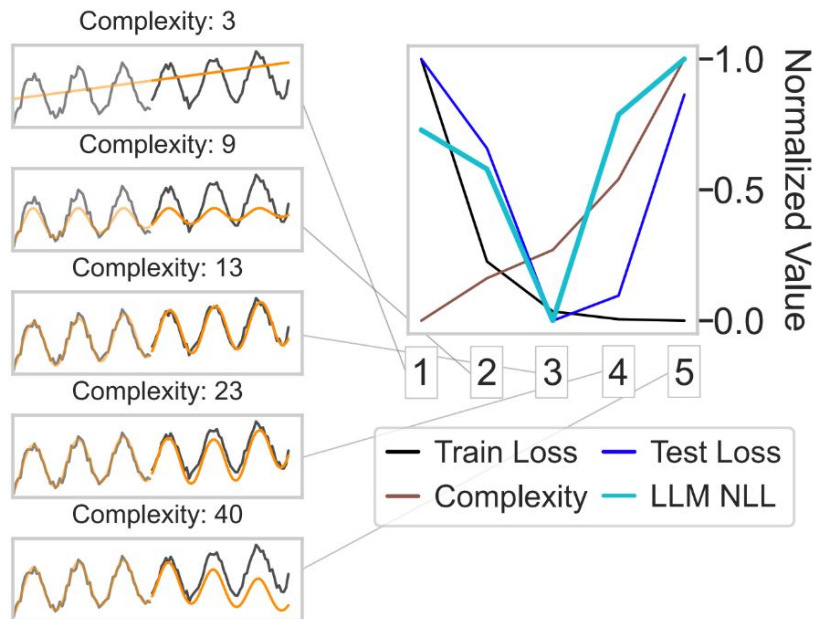
Pushing via Clicker Training



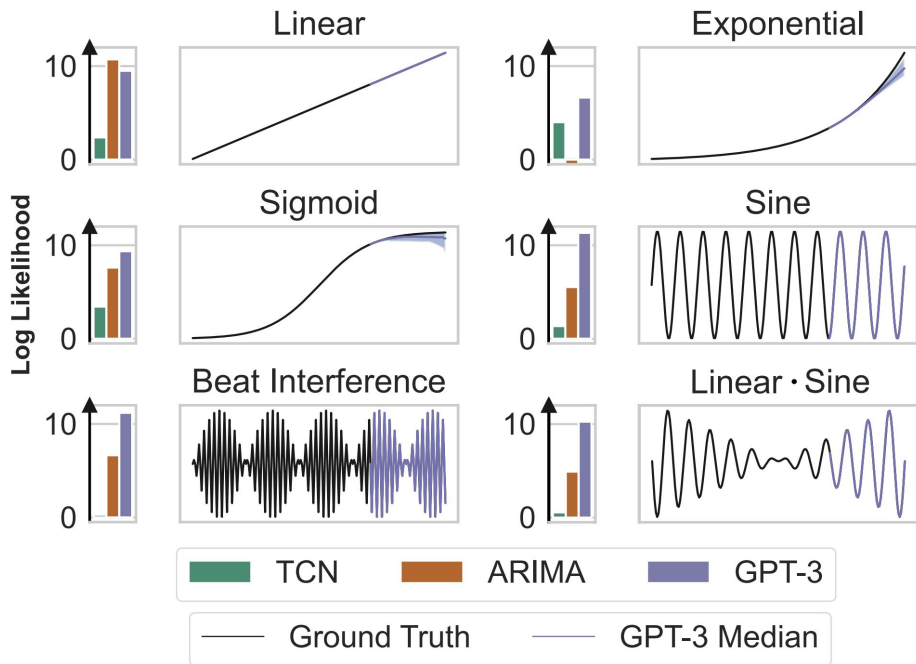
Why Do LLMs Work?



Why Do LLMs Work?



Why Do LLMs Work?



Outline

- ❑ Motivating large language models (LLMs) for forecasting
- ❑ Our method (LLMTime)
- ❑ Experimental validation
- ❑ Understand why LLMs work for forecasting
 - ❑ **Future work**

Future Work

- Vision-language models (VLMs) show that many different sources can be combined as input with fine-tuning.



Liu et al. 2023

User Please read the text in this image and return the information in the following JSON format (note xxx is placeholder, if the information is not available in the image, put "N/A" instead).

```
{ "class": xxx, "DLN": xxx, "DOB": xxx, "Name": xxx, "Address": xxx, "EXP": xxx, "ISS": xxx, "SEX": xxx, "HGT": xxx, "WGT": xxx, "EYES": xxx, "HAIR": xxx, "DONOR": xxx }
```

GPT-4V

```
{ "class": "C", "DLN": "1234568", "DOB": "08/31/1977", "Name": "Ima Cardholder", "Address": "2570 24th Street, Anytown, CA 95818", "EXP": "08/31/2014", "ISS": "08/31/2011", "SEX": "F", "HGT": "5-05", "WGT": "125 lbs", "EYES": "BRN", "HAIR": "BRN", "DONOR": "VETERAN" }
```

LLaVA

```
{ "class": "Driver License", "DLN": "CA123 456 789", "DOB": "01/01/1990", "Name": "Amy", "Address": "Anytown, CA", "EXP": "08/31/2023", "ISS": "Veteran", "SEX": "F", "HGT": "5'5", "WGT": "130", "EYES": "Brown", "HAIR": "Brown", "DONOR": "Unknown" }
```

LLaVA-1.5

```
{ "class": "C", "DLN": "1132456789", "DOB": "08/23/1971", "Name": "Ima Cardholder", "Address": "Anytown, State, Zip", "EXP": "08/23/2014", "ISS": "California", "SEX": "F", "HGT": "5'5", "WGT": "123", "EYES": "Brown", "HAIR": "Brown", "DONOR": "N/A" }
```

Future Work

❑ Vision-language models (VLMs) show that many different sources can be combined as input with fine-tuning.

❑ The real challenge in most forecasting is combining many source of information to make a good prediction.



Liu et al. 2023

User Please read the text in this image and return the information in the following JSON format (note xxx is placeholder, if the information is not available in the image, put "N/A" instead).

```
{ "class": xxx, "DLN": xxx, "DOB": xxx, "Name": xxx, "Address": xxx, "EXP": xxx, "ISS": xxx, "SEX": xxx, "HGT": xxx, "WGT": xxx, "EYES": xxx, "HAIR": xxx, "DONOR": xxx }
```

GPT-4V

```
{ "class": "C", "DLN": "1234568", "DOB": "08/31/1977", "Name": "Ima Cardholder", "Address": "2570 24th Street, Anytown, CA 95818", "EXP": "08/31/2014", "ISS": "08/31/2011", "SEX": "F", "HGT": "5-05", "WGT": "125 lbs", "EYES": "BRN", "HAIR": "BRN", "DONOR": "VETERAN" }
```

LLaVA

```
{ "class": "Driver License", "DLN": "CA123 456 789", "DOB": "01/01/1990", "Name": "Amy", "Address": "Anytown, CA", "EXP": "08/31/2023", "ISS": "Veteran", "SEX": "F", "HGT": "5'5", "WGT": "130", "EYES": "Brown", "HAIR": "Brown", "DONOR": "Unknown" }
```

LLaVA-1.5

```
{ "class": "C", "DLN": "1132456789", "DOB": "08/23/1971", "Name": "Ima Cardholder", "Address": "Anytown, State, Zip", "EXP": "08/23/2014", "ISS": "California", "SEX": "F", "HGT": "5'5", "WGT": "123", "EYES": "Brown", "HAIR": "Brown", "DONOR": "N/A" }
```
