TsLLM: Augmenting LLMs for General Time Series Understanding and Prediction

Felix Parker, Nimeesha Chan, Chi Zhang, Kimia Ghobadi

Center for Systems Science and Engineering, Johns Hopkins University

JOHNS HOPKINS UNIVERSITY

Limitations of Conventional Time Series Methods

1. Can’t incorporate other data modalities, contextual information, and domain knowledge

2. Inflexible, require re-training or new analysis

Solution: Multimodal LLMs

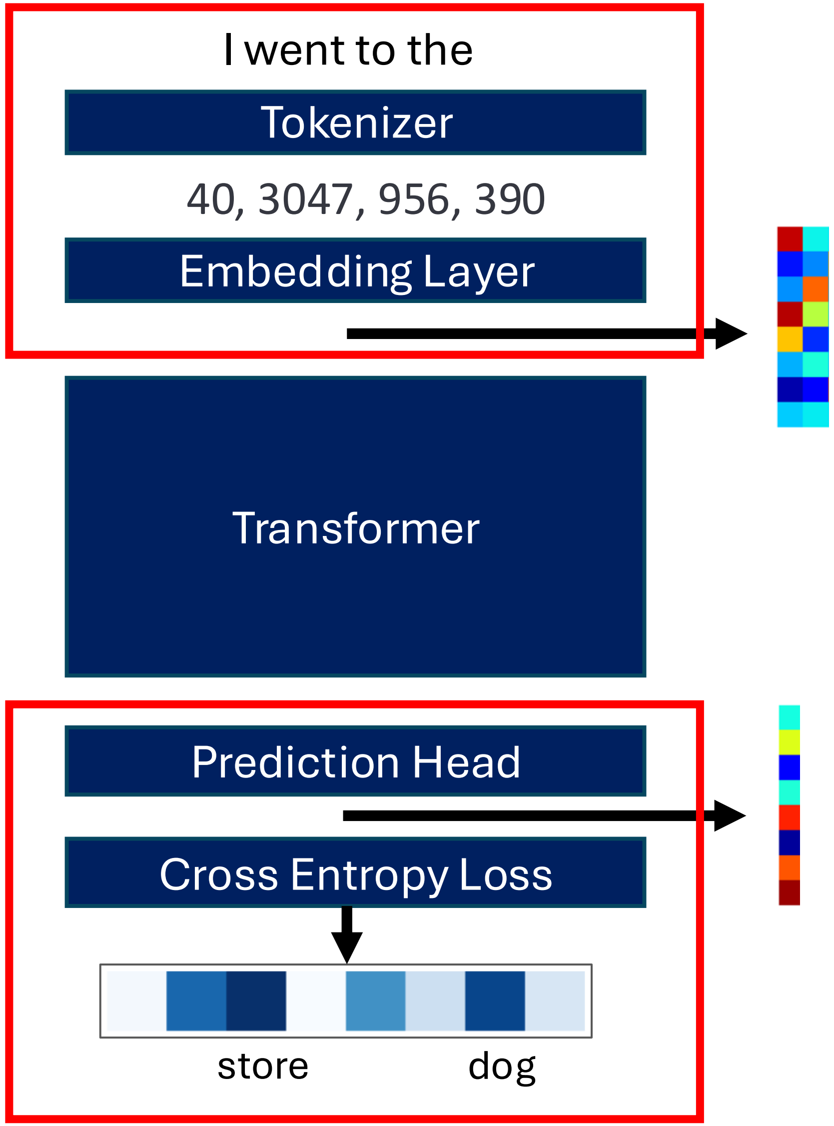
- LLMs are powerful models with general knowledge and skills
- LLMs can integrate information across modalities
- Strong reasoning abilities
- An intuitive interface for users with varying skills and backgrounds

The challenge?

- Struggle with analyzing high-volume numerical data, *particularly high-frequency time series data*

Why can't LLMs already do this?

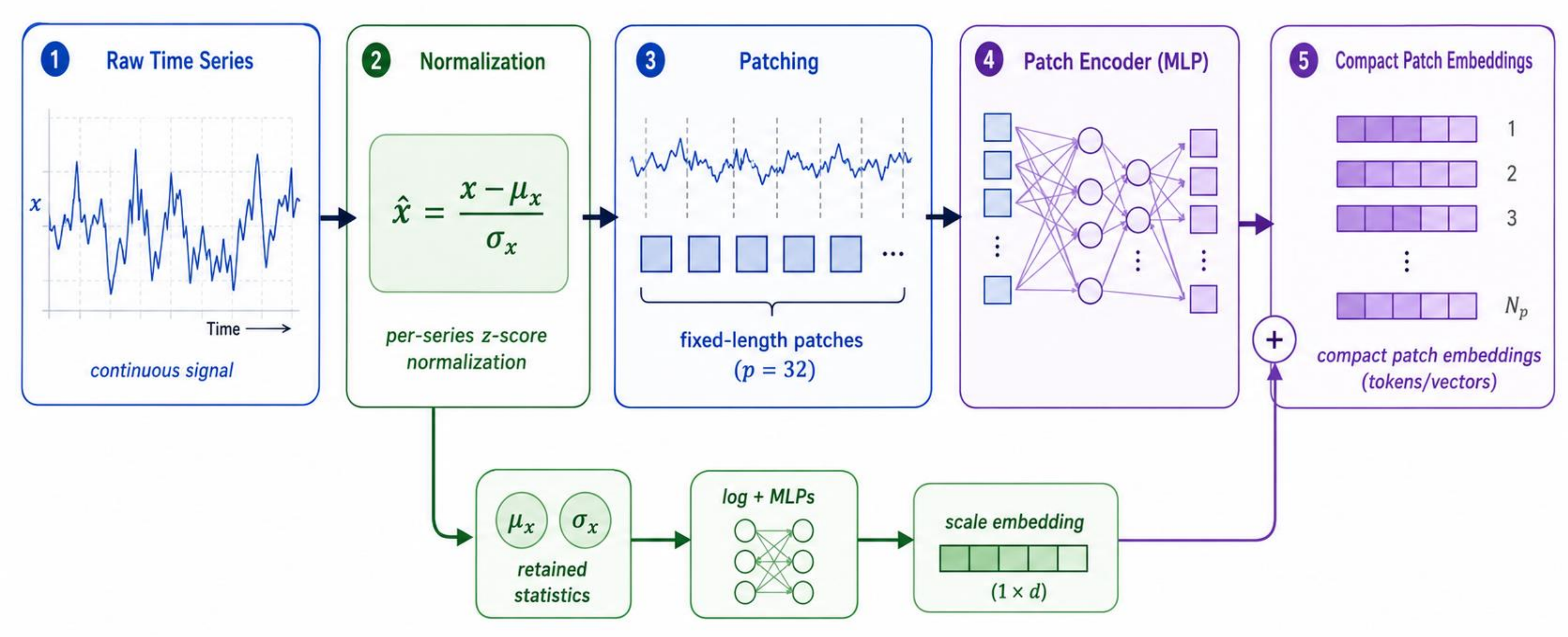
- Tokenization doesn't work well for numerical data
  - Splits up long numbers, requires many tokens, low information density
  - LLMs are trained with a classification loss that's inefficient for numbers
- Plot encodings limit understanding and flexibility
- LLMs are not trained for time series understanding
  - Very different than modeling natural language
  - Most numerical data is filtered out of training datasets – less than 1 part per million remaining



Goals

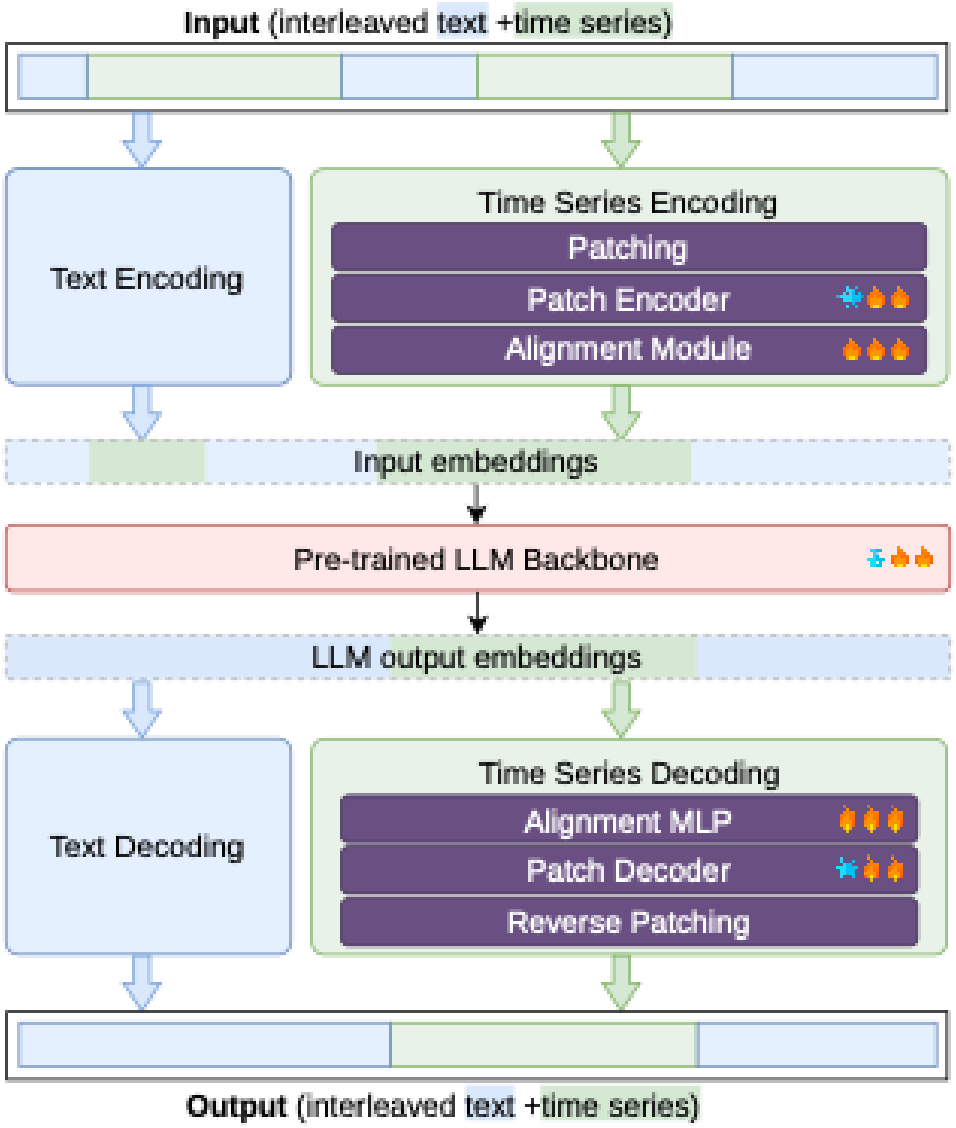
- Teach an LLM to understand numerical time series + text data
- Solve traditional time series tasks with contextual information
  - Forecasting, classification, etc.
  - Context from unstructured text, tabular data, imaging, etc.
- Solve novel time series tasks that heavily involve text
  - Question-answering, report generation
- Build a foundation model
  - Doesn’t require re-training for every dataset
  - Can do in-context learning to adapt to new data

Time Series Codec



LLM Integration

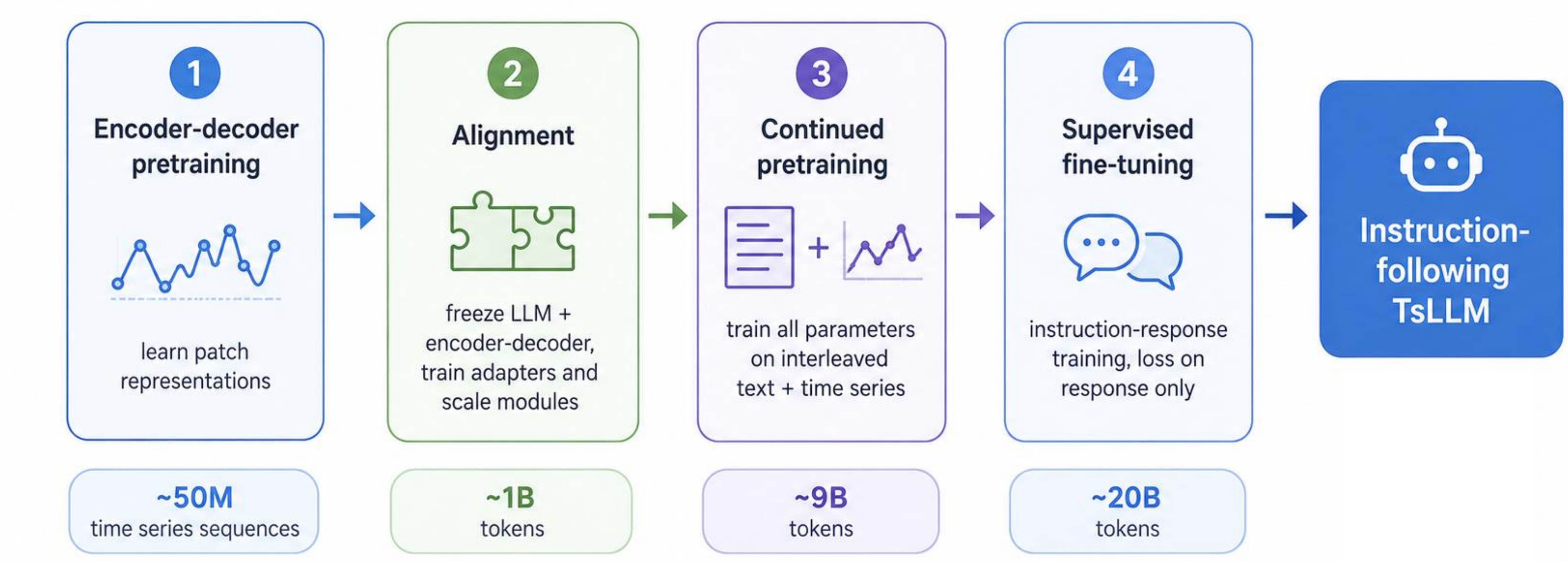
- Pretrained LLM backbone
- Time series encoder + decoder
- Alignment modules
- Can interleave text and time series in inputs and outputs



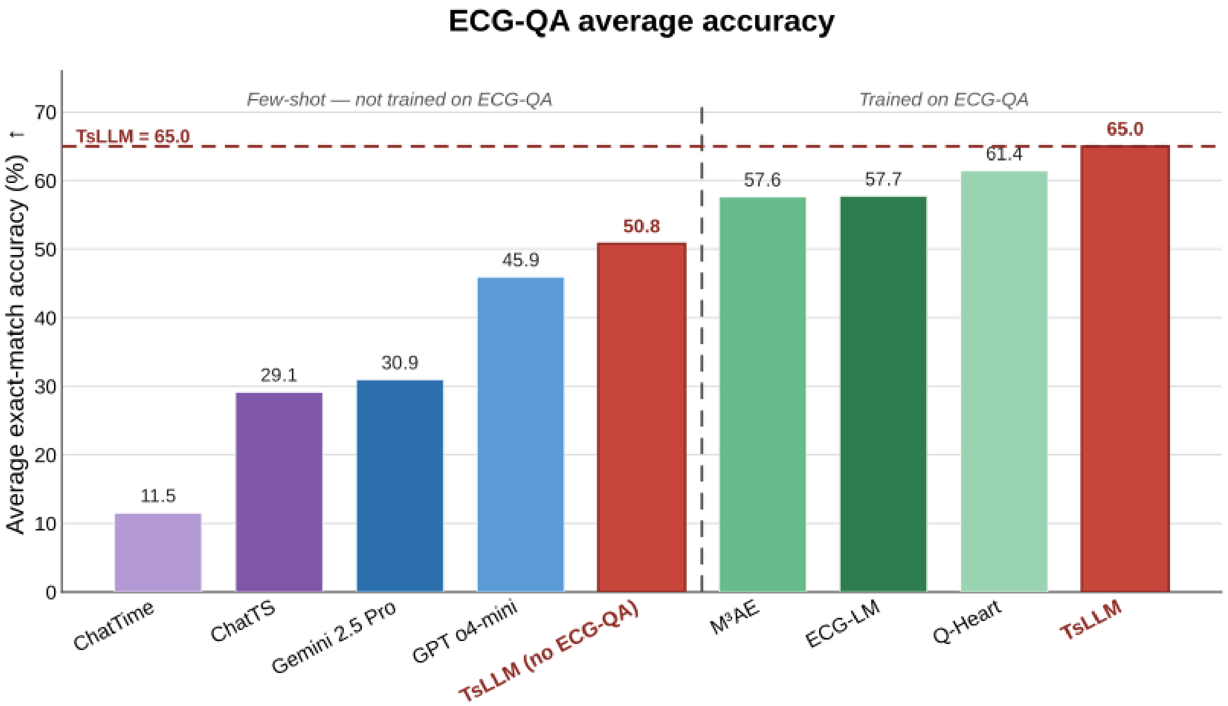
Data Collection

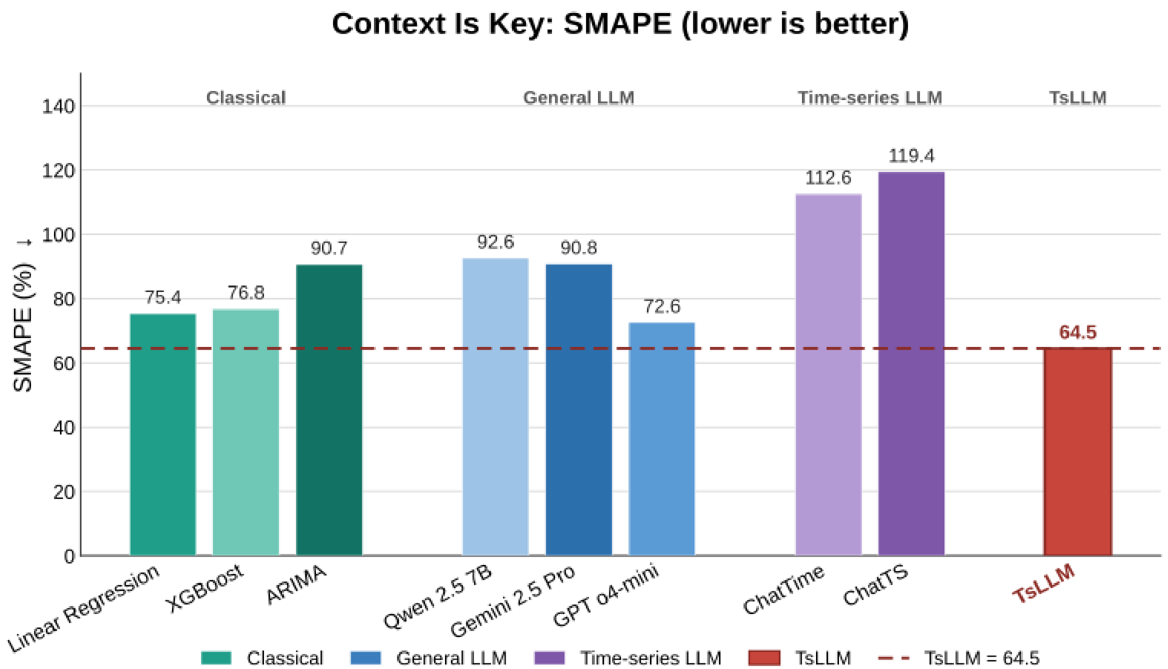
- LLMs are very data hungry!
- No existing large-scale datasets of paired time series + text
- Major effort to collect sufficient training data
- Sources:
  - Some small datasets with paired time series + text
  - Large time series datasets + converted metadata/tabular data into text
  - Combined time series dataset with separate relevant text source
  - Augmented data using existing LLMs with guidance and filtering
- Total: ~27 billion tokens

Training Procedure



Results





ECG-QA: clinical question-answering on electrocardiogram data

Context Is Key: forecasting tasks across domains that require using natural-language context

Conclusions

- Extremely flexible
- Effectively uses text-based context and domain knowledge
- Can perform non-standard time series tasks with very strong performance
  - Question-answering, contextual forecasting
- Does not require re-training for new tasks
  - Can adapt with in-context examples

TsLLM: Augmenting LLMs for General Time Series Understanding and Prediction

Felix Parker, Nimeesha Chan, Chi Zhang, Kimia Ghobadi

Center for Systems Science and Engineering, Johns Hopkins University

TsLLM: