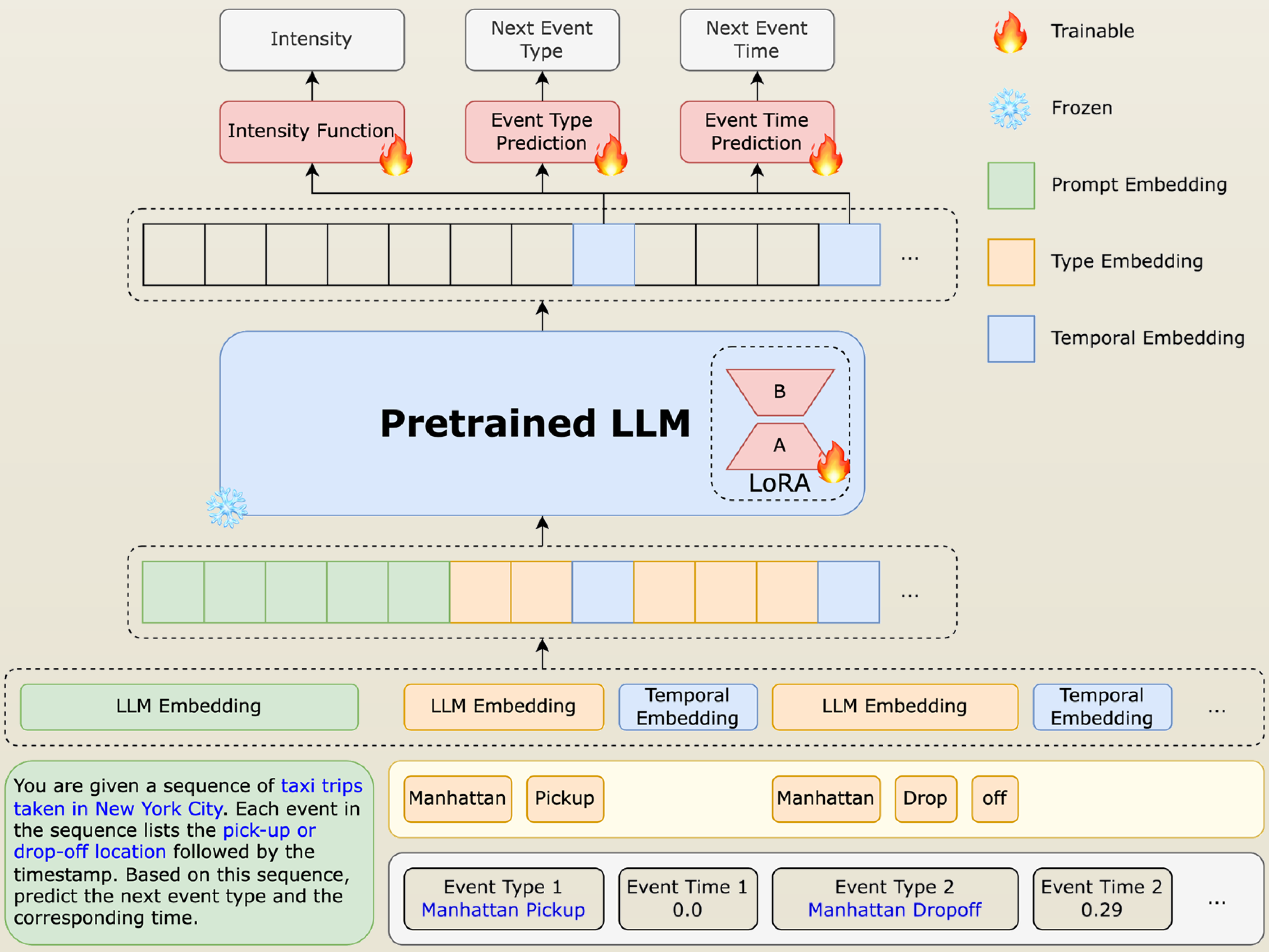
TPP-LLM: Modeling Temporal Point Processes by Efficiently Fine-Tuning Large Language Models

Zefang Liu, Yinzhu Quan (Georgia Institute of Technology)

Contact: liuzefang@gatech.edu

Introduction

- **Temporal Point Processes (TPPs)** are essential for **modeling sequences of events over time** in domains such as social networks, urban mobility, and e-commerce.
- Traditional TPP models struggle to capture both **semantic richness** and **complex temporal patterns**, often relying on categorical event representations and handcrafted features.
- **TPP-LLM** introduces a novel integration of **Large Language Models (LLMs)** with **TPPs**, enabling semantic-aware and temporally-informed event prediction.



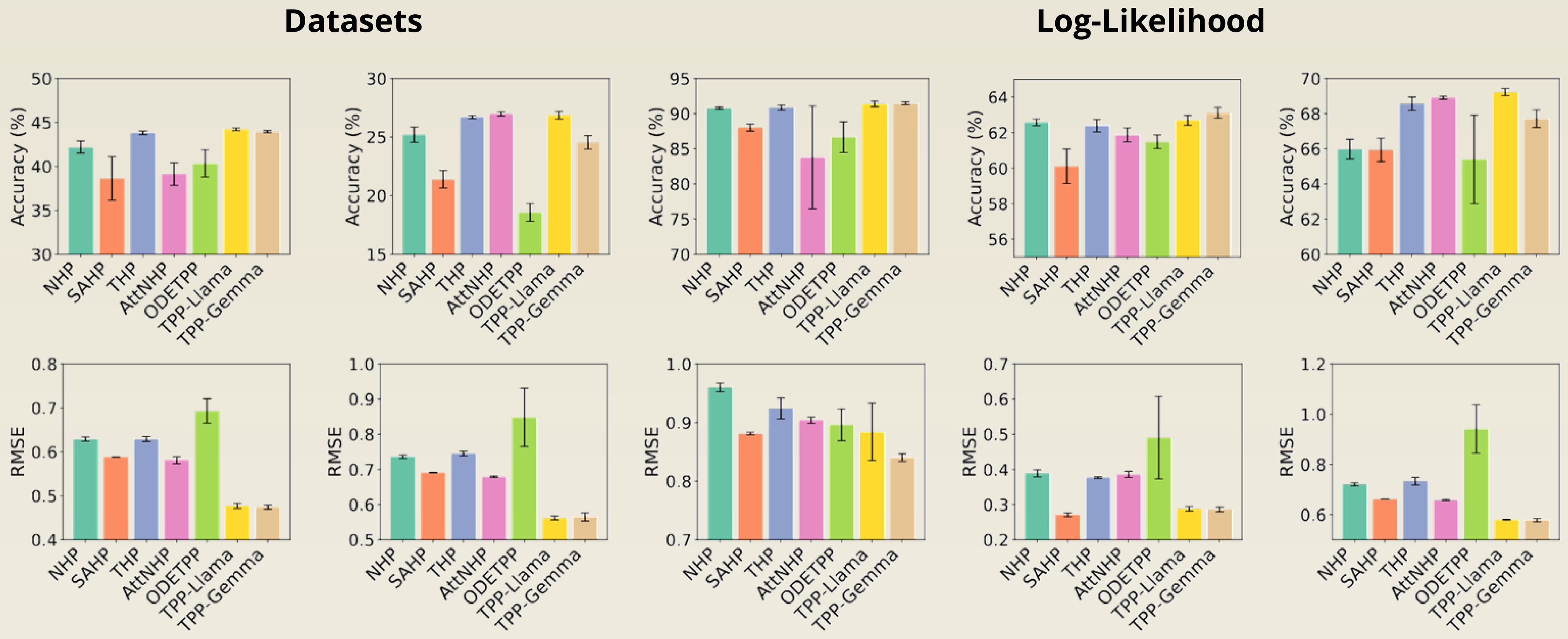
Prompt **Event Sequence**

Dataset	# of Types	# of Events	# of Seq.
Stack Overflow	25	187,836	3,336
Chicago Crime	20	202,333	4,033
NYC Taxi Trip	8	362,374	2,957
U.S. Earthquake	3	29,521	3,009
Amazon Review	18	127,054	2,245

Datasets

Model	StackOverflow	Crime	Taxi	Earthquake	Amazon
NHP	-2.005	-2.604	0.366	-0.450	-1.196
SAHP	-6.320	-6.069	-0.228	**0.193**	-4.201
THP	-1.877	-2.493	0.217	-0.513	-1.083
AttNHP	-1.798	**-2.432**	**0.446**	-0.481	**-0.959**
ODETPP	-2.402	-4.152	-0.450	-0.511	-1.808
TPP-Llama	**-1.777**	-2.451	0.271	-0.475	-1.011
TPP-Gemma	-1.785	-2.480	0.332	-0.479	-1.075

Log-Likelihood



Stack Overflow **Chicago Crime** **NYC Taxi Trip** **U.S. Earthquake** **Amazon Review**

Experiments

- Evaluated on **five real-world datasets**: Stack Overflow, Chicago Crime, NYC Taxi, U.S. Earthquake, and Amazon Review, each with event timestamps and textual type descriptions.
- Compared with **neural TPP baselines**: NHP, SAHP, THP, AttNHP, and ODETPP, using **log-likelihood**, type prediction **accuracy**, and time prediction **RMSE**.
- Built on lightweight foundation models, including **TinyLlama-1.1B** and **Gemma-2B**, using **4-bit quantization** and **LoRA** for efficient adaptation.
- **TPP-LLM** consistently outperforms most baselines, especially in event type accuracy and time prediction on complex or semantically rich sequences.

Preliminaries

- **Marked TPPs** model event sequences $\mathcal{S} = \{(t_1, k_1), , \ldots, (t_n, k_n)\}$, where each event has a timestamp t_i and type k_i. The goal is to predict the next event's time and type given the history $\mathcal{H}_t$.
- The **conditional intensity function** $\lambda(t, k|\mathcal{H}_t)$ defines the instantaneous rate of observing an event of type k at time t:

$$\lambda(t, k|\mathcal{H}_t) = \lim_{\Delta t \to 0} \frac{\mathbb{E}[N_k(t + \Delta t) - N_k(t)|\mathcal{H}_t]}{\Delta t}.$$

- **Neural TPPs** use models like RNNs or transformers to learn the intensity function from data. Given event embeddings e_i, hidden states are updated via $h_i = f(h_{i-1}, e_i)$ to capture complex temporal and type dependencies.

Methodology

- **TPP-LLM** models event sequences by combining **textual descriptions** of event types with **temporal embeddings**, enabling the model to learn both semantic and temporal patterns.
- Each event type k_i is tokenized and embedded using a **pretrained LLM**, while the corresponding time t_i is mapped to a **temporal embedding** $f_{\text{temporal}}(t_i)$, such as sinusoidal positional encoding.
- The combined embeddings are processed by a **decoder-only transformer**, generating hidden states h_i used to compute:
  - The **intensity function**: $\lambda_k(t|\mathcal{H}_t) = \text{softplus}\big(\alpha_k(t - t_i) + \boldsymbol{w}_k^T \boldsymbol{h}_i + b_k\big)$.
  - The **next event type**: $\hat{k}_{i+1} = \text{argmax}(\boldsymbol{W}_{\text{type}}^T \boldsymbol{h}_i + \boldsymbol{b}_{\text{type}})$.
  - The **next event time**: $\hat{t}_{i+1} = \boldsymbol{w}_{\text{time}}^T \boldsymbol{h}_i + b_{\text{time}}$
- **LoRA-based fine-tuning** is applied to adapt the LLM efficiently by injecting trainable low-rank matrices into attention layers, reducing the number of parameters while preserving performance.

Conclusion

- **TPP-LLM** integrates pretrained LLMs with TPPs, enabling joint modeling of **event semantics** and **temporal dynamics** for improved **event prediction**.
- Through **PEFT**, the model achieves strong performance across diverse **real-world datasets**, consistently outperforming existing baselines in both **sequence modeling** and **event prediction**.

Paper **Data** **Code**

References

- Mei, Hongyuan, and Jason M. Eisner. "The neural Hawkes process: A neurally self-modulating multivariate point process." *Advances in neural information processing systems* 30 (2017).
- Zhang, Qiang, et al. "Self-attentive Hawkes process." *International conference on machine learning*. PMLR, 2020.
- Zuo, Simiao, et al. "Transformer Hawkes process." *International conference on machine learning*. PMLR, 2020.
- Xue, Siqiao, et al. "EasyTPP: Towards open benchmarking temporal point processes." *arXiv preprint arXiv:2307.08097* (2023).