Dec 5

Brock University

Knowledge Graphs for Trace Analysis

Alireza Ezaz
Naser Ezzati-Jivan

Key Challenges in Tracing

- Extensive amount of data
- Difficulty in comprehending trace data
- **Example:**
  - Millions of records make it hard to find specific interactions (e.g., threads working with file descriptor ID 24)

Our goal is to develop methods that help us navigate this data efficiently, so we can answer specific questions about the system without getting overwhelmed.

Step 1: Generating Translations of Traces

- **Our Initial Approach:** Created human-readable translations by incorporating expert feedback
- Developed specific functions for each kernel event

Kernel Event Type	Tailored function to Translate
sched_wakeup	def translate_sched_wakeup(self, row)
syscall_entry_ioctl	def translate_syscall_entry_ioctl(self, row):
syscall_entry_open	def translate_syscall_entry_open(self, row):
syscall_entry_socket	def translate_syscall_entry_socket(self, row):
sched_switch	def translate_sched_switch(self, row):

Generating Translations of Traces

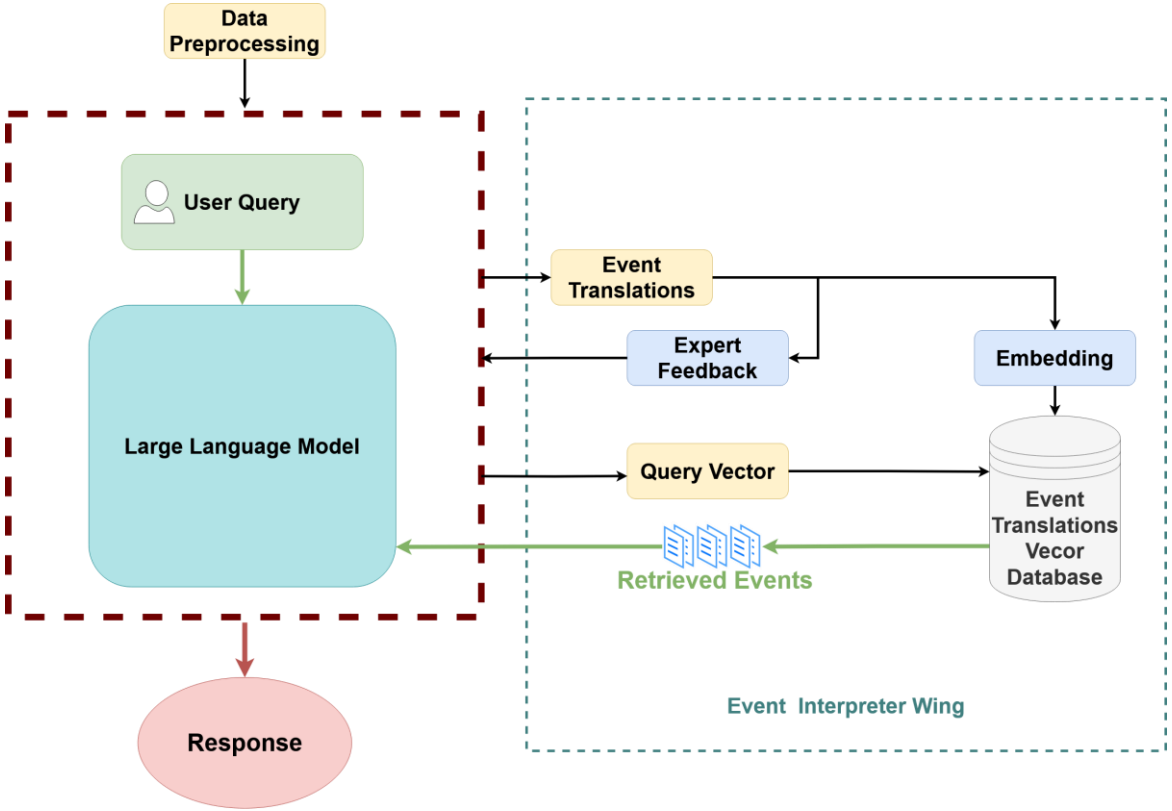
Raw Trace Data

09:32:47.799 450
353,kernel_3,3,sched_switch,prev_comm=swapper/3, prev_tid=0, prev_prio=20, prev_state=0, next_comm=lttng-consumerd, next_tid=2208, next_prio=20, context.packet_seq_num=0, context.cpu_id=3,[magic=3254525889, uuid=[82, 247, 60, 150, 234, 196, 120, 73, 134, 56, 235, 15, 159, 101, 123, 78], stream_id=0, stream_instance_id=3],[timestamp_begin=1127621697314104, timestamp_end=1127651320322338, content_size=33428696, packet_size=33456128, packet_seq_num=0, events_discarded=0, cpu_id=3],nan,nan,nan,nan,nan,nan

Human-Readable Translation

At 09:32:47.799 450 353, CPU 3 switched from task swapper/3 (TID 0, priority 20, state 0) to task lttng-consumerd (TID 2208, priority 20).

Answering System Questions with Event Translations and LLMs



Limitations of the Initial Approach

1. Generating human-readable translations made the data more accessible
2. Performs well in some categories of questions like event sequences

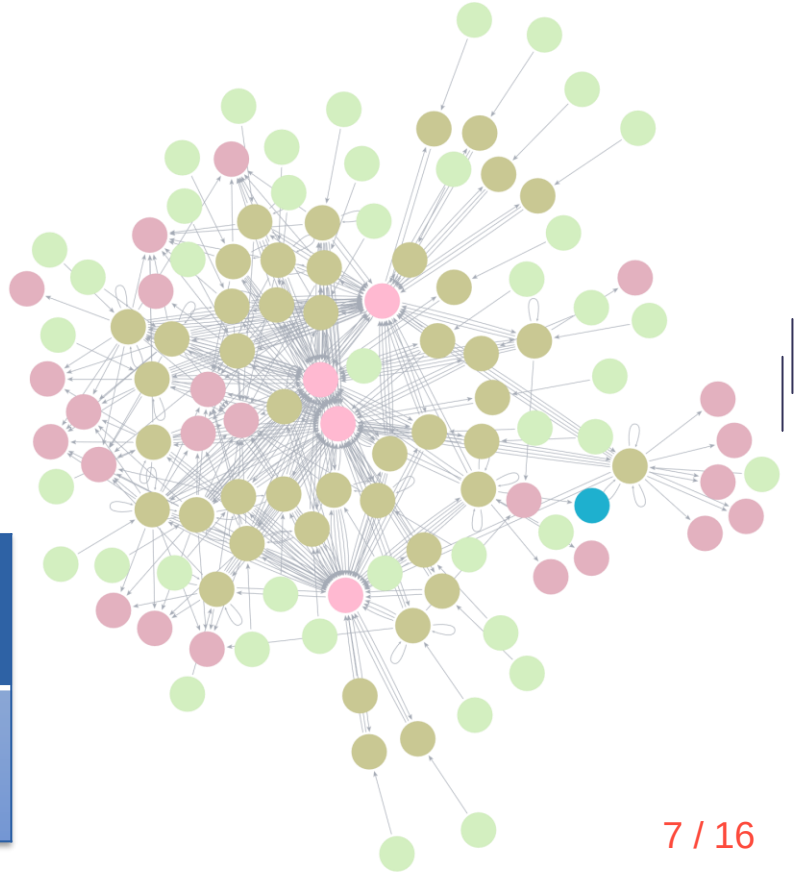
- **Examples it helps:**
  - What events occurred before Thread T_5123 switched out from CPU 3?
  - What is the sequence of events for Process P_5124?

1. May not capture structural data or dependencies
2. might need to retrieve and process a huge number of event translations

- **Examples it fails:**
  - Which files is CPU 3 interacting with?
  - Which threads are interacting with CPU 0?
  - Which CPUs are Threads T_2208 and T_5123 running on?

Step 2: Development of a Knowledge Graph

- **Step 2:** Created a Knowledge Graph to capture structural insights and dependencies
- Developed specific functions for each kernel event to extract entities and relationships from raw trace data



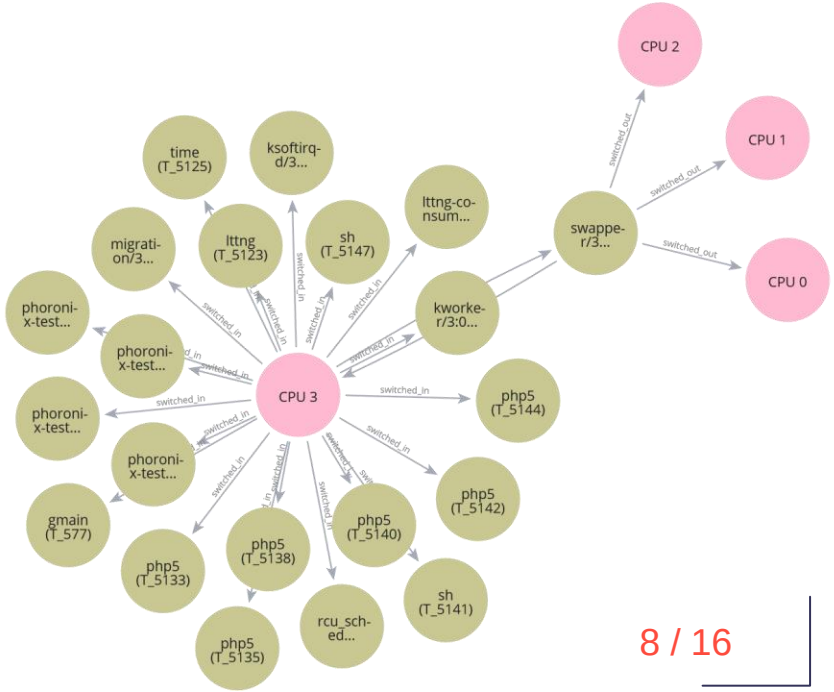
Kernel Event Type	Tailored function to Extract Entities and Relations
sched_wakeup	def handle_sched_wakeup(self, row)

Knowledge Graph Characteristics

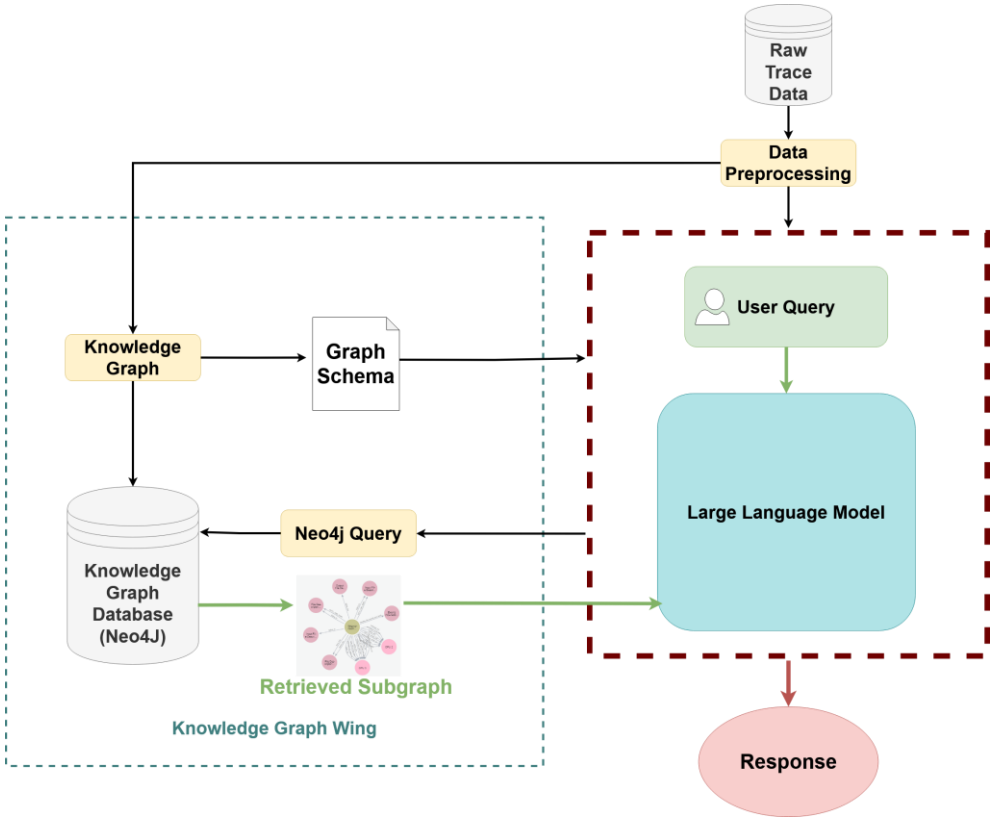
1. Directed graph structure
2. Weighted graph structure
3. Event-specific entity and relation extraction
4. Multiple relationships between the same pair of entities
5. Event-based node and edge properties

Nodes: CPUs, Processes, Threads, Files, Network Sockets, ...

Edges: Writes to, Reads from, wakes up, ...



Answering System Questions with Knowledge Graph and LLMs



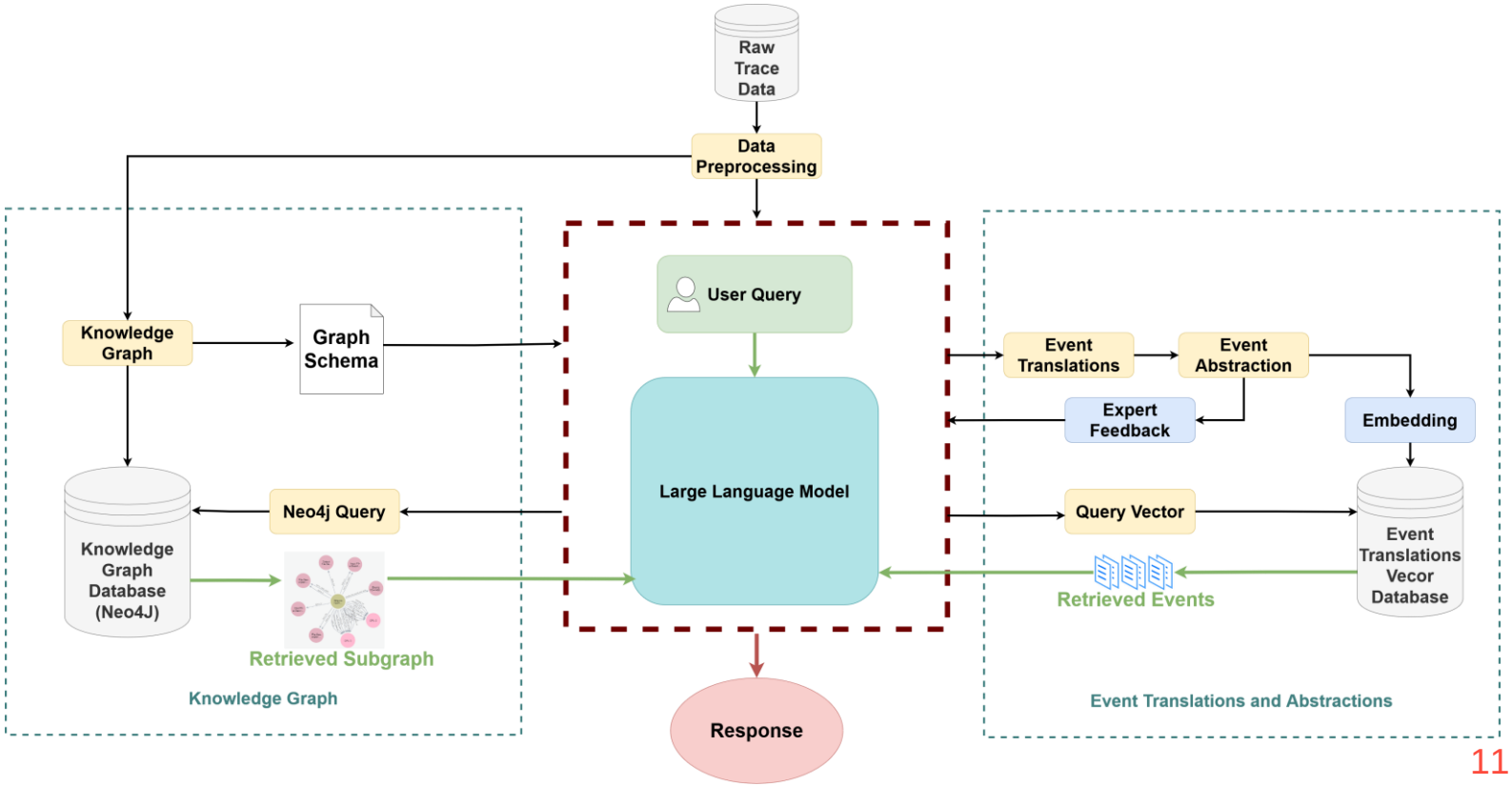
Querying the Knowledge Graph

- **Advancements**
  - Initially performed manual queries
  - Stored the graph in Neo4j for efficient querying
  - Used an LLM with prompt Engineering to generate accurate Cypher queries

- **Benefits:**
  - Captures explicit entities and relationships
  - Improves accuracy on dependency-related questions

Example Cypher Query MATCH (t:Thread)-[r]->(f:File {fd: '24'})RETURN DISTINCT t

Answering System Questions with Knowledge Graph, Translations and LLMs



We Still have challenges working with large amounts of data

Focusing on Knowledge Graphs and Machine Learning Models

- Our Current Focus:
  - Solely leveraging knowledge graph
  - Testing different machine learning models

- Starting Point:
  - Graph Neural Networks (GNNs) for effective handling of large datasets and capturing structural data
  - Question Answering using GNN
    - Ex: `Node Classification for Direct Answers`
      `If the answer to a question corresponds to a specific node in the KG (e.g., “Which file was accessed by process X?”), a GNN can classify nodes to determine the most likely candidate.`

Current Work and Future Directions

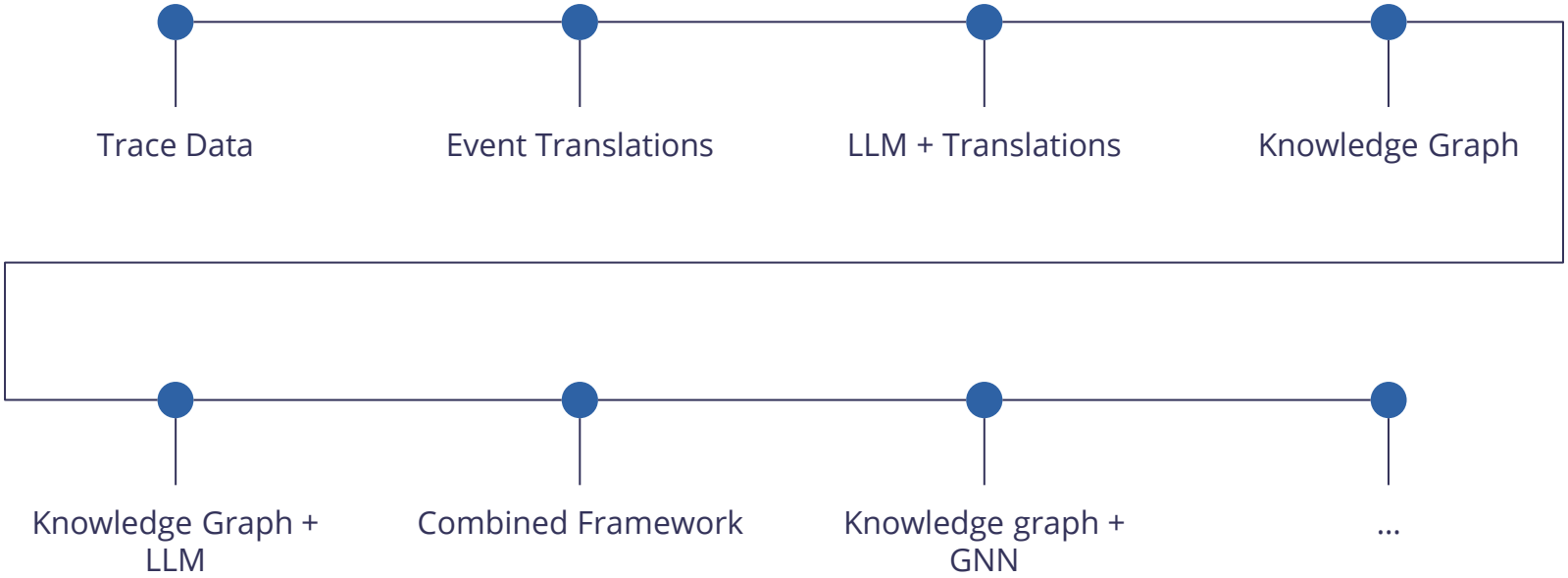
1. Exploring Graph Transformers for further improvements

2. Creating knowledge graphs from different aspects like security, performance and ...

3. Focusing on different types of questions
    1. Single-Hop Queries
    2. Multi-Hop Queries: Retrieve indirect relationships involving multiple edges
    3. Path Queries
    4. Reachability Queries
    5. Subgraph Queries

4. Using multiple knowledge graphs (Maybe temporal)
5. ...

Summary of Contributions



Thanks!

Do you have any questions?

sezaz@brocku.ca

www.linkedin.com/in/s-alireza-ezaz

BTW, I am Looking for summer internships.
Let's discuss if you know any opportunities ☺