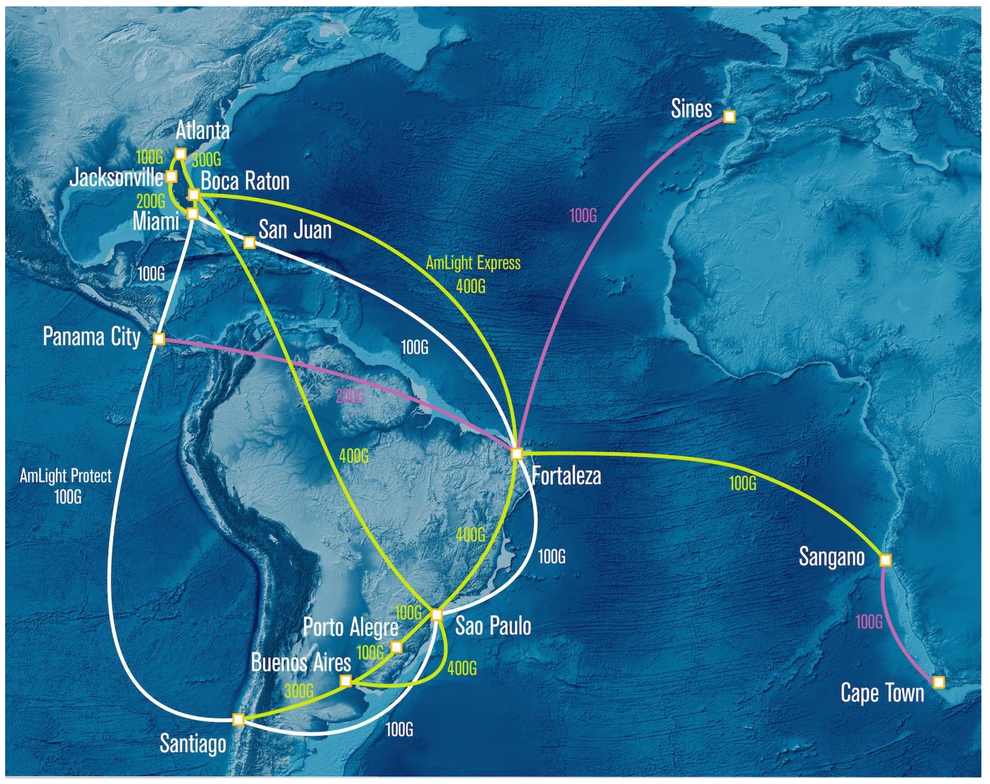
LLMDaL: LLM-Driven Data Labeling for Training ML Models

Kemal Akkaya[1], Julio Ibarra[2], Yanzhao Wu[2], Jeronimo Bezerra[2], Abdulhadi Sahin[2],
Virginia Commonwealth University[1], Florida International University[2]

https://adwise-vcu.github.io/adwise-lab/projects/llm_data_labeling.html

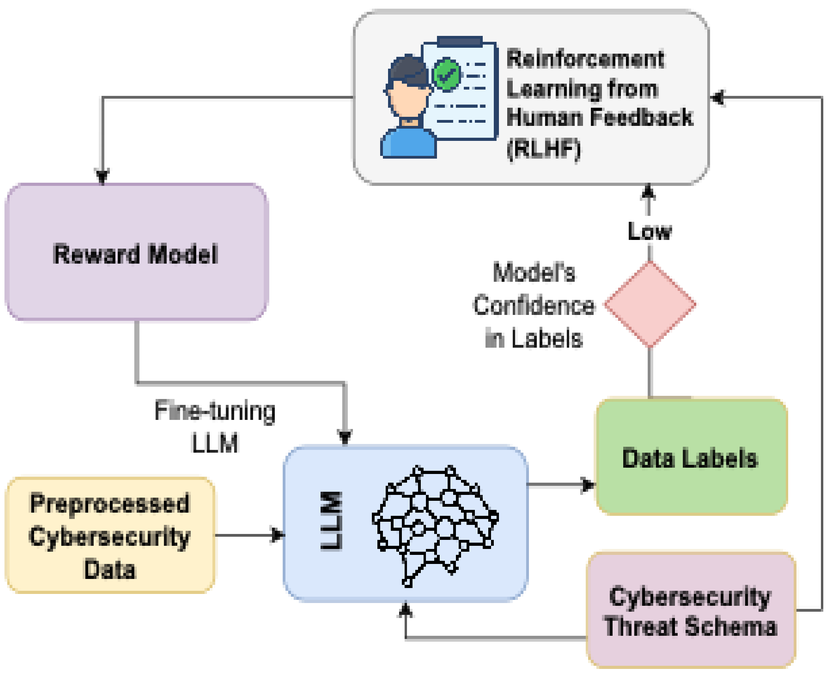
Research Impact and Key Takeaways

- Bridges the gap between **production data and AI-based** solutions.
- Enables **scalable, automated LLM-driven labeling** for cybersecurity datasets
- Reduces reliance on **manual expert labeling**
- Generalizable to **other domains beyond cybersecurity**



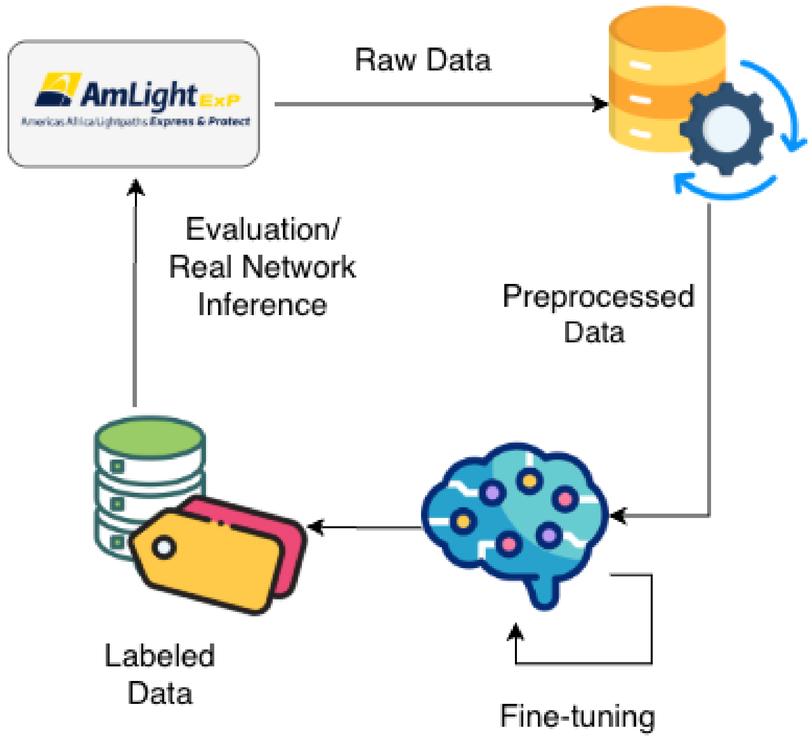
Scientific Cybersecurity Dataset Needs and Gaps Addressed

- Scientific cyberinfrastructures are both critical and vulnerable and they operate under unique high-throughput and low-latency.
- Current AI/ML approaches face real-world limitations
- Progress is constrained by data challenges such as privacy, legal, and operational barriers.
- The project addresses these gaps by delivering high-quality labeled cybersecurity datasets from real-world traffic on AmLight (Americas Lightpaths) by utilizing an LLM-driven labeling agent for automated data curation and labeling.



Generating the Cybersecurity Dataset

- Packet-level INT and network logs; IDS alerts, simulated attacks
- In-house parsers for bidirectional flows with entropy features
- LLM fine-tuned on cybersecurity knowledge and network data
- LLM labeling agent with self-refinement, ensemble methods, and expert verification



Benefits to Scientific Cyberinfrastructure

- Production-grade, packet-level labeled datasets with multilevel classifications
- LLM-driven labeling agent for automatic, high-quality label generation
- Enhances existing cybersecurity solutions and enables new approaches for cyberinfrastructure

Result Dissemination Plans

- Dataset available via LLMDaL website, updated annually to reflect evolving threats
- Code, tools, and resources open-sourced on GitHub
- Tuning approaches published at scientific conferences

Risks Versus Potential For Advances

- Complex multi-layer TCP/IP data processing required for LLM training datasets
- Consistent labeling needed to prevent hallucination and concept drift
- Opportunity for scalable, automated LLM-driven labeling, applicable to other domains