HuggingGraph: Understanding the Supply Chain of LLM Ecosystem

Mohammad Shahedur Rahman

Peng Gao

Yuede Ji

Large Language Model (LLM)

- LLM refers to
  - A specialized type of AI model trained on a large amount of data, to understand existing content and generate human-like language. [NIST Glossary, CSRC (Computer Security Resource Center)].

- Example: ChatGPT, Llama, etc.

LLM Hosting Platforms

- Hosting platforms democratize model development and deployment.
  - E.g., Hugging Face, PyTorch Hub, ONNX.

- E.g., Hugging Face hosted over 2M models and 500K datasets, and its rapidly growing.

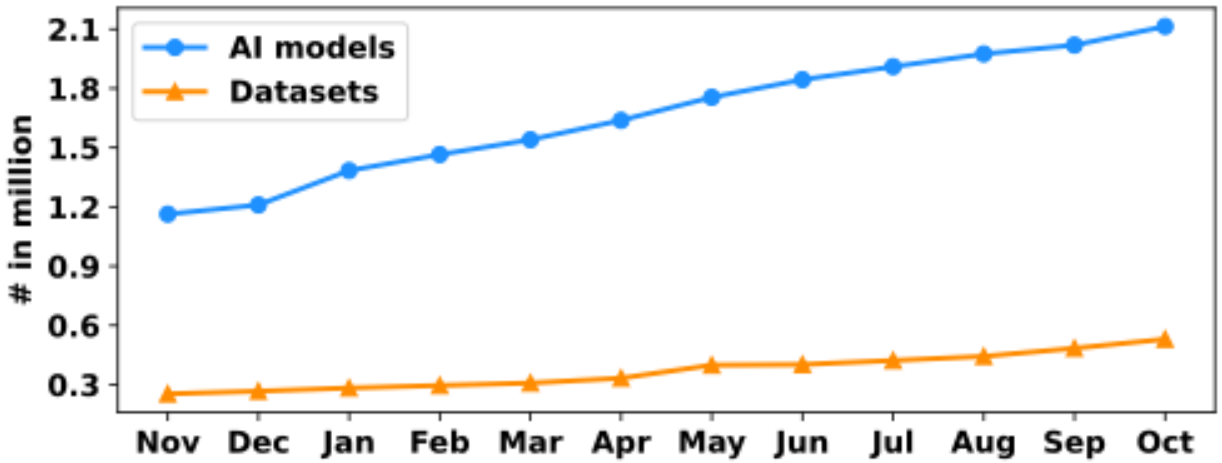
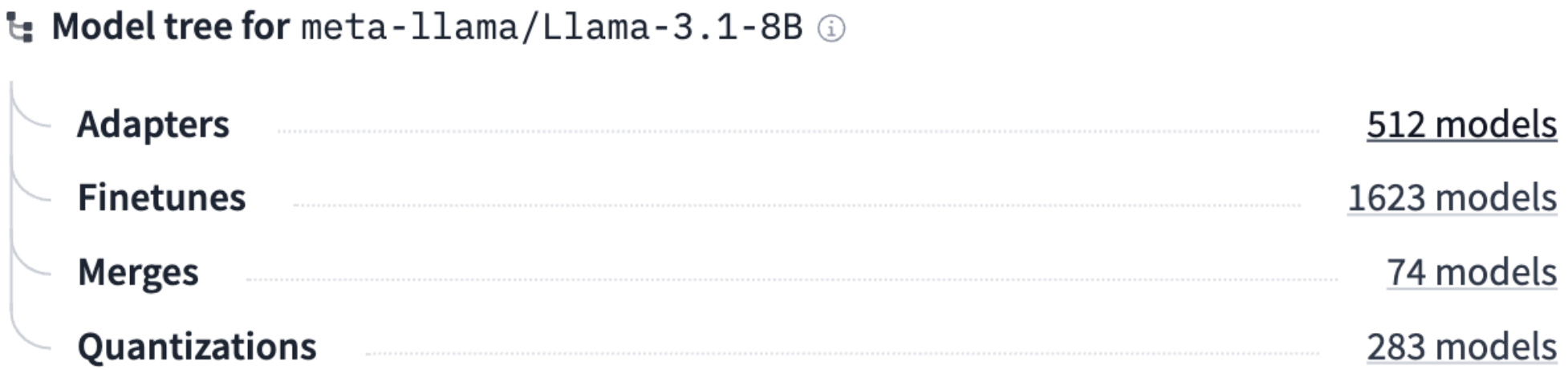


Fig: AI models and datasets (in millions) on *Hugging Face* from November 2024 to October 2025.

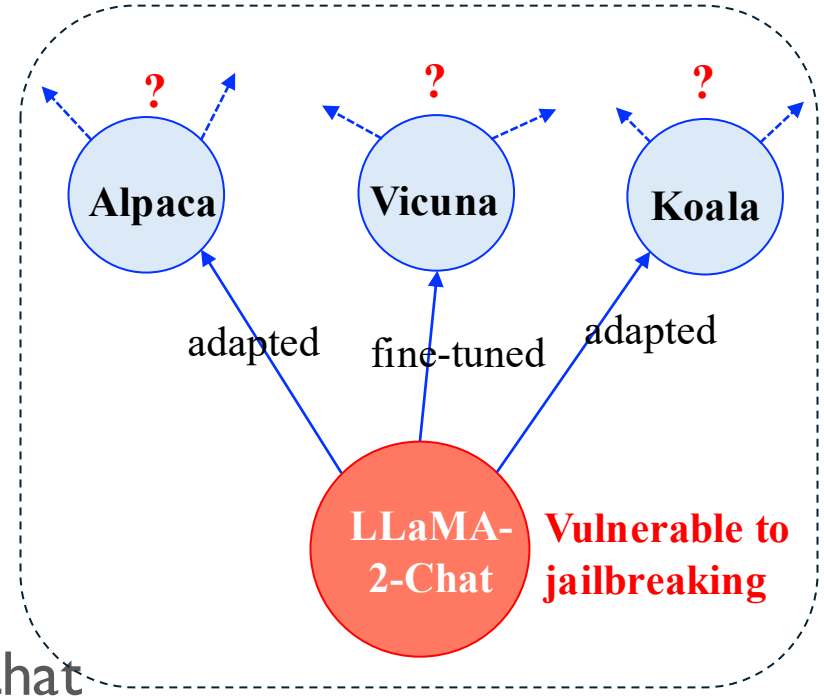
What is LLM Supply Chain?

- LLM supply chain refers to *the interconnected lifecycle of models, datasets, software libraries, computing, and other relevant components.*
  - Across stages of training, fine-tuning, adaptation, and deployment.
  - Example
    - Model tree for *meta-llama/Llama-3.1-8B* from Hugging Face



Why Do We Care about LLM Supply Chain?

- Motivation:
  - Security assurance: Identify *inherited vulnerabilities, biases and malicious components*, that can propagate through LLM supply chain.
  - Dataset origin: Ensures transparency of data sources, verifying reliability, fairness, and legal compliance.
  - Model evolution: Traces relationships among models that reveals how LLMs evolve from base to task-specific variants.



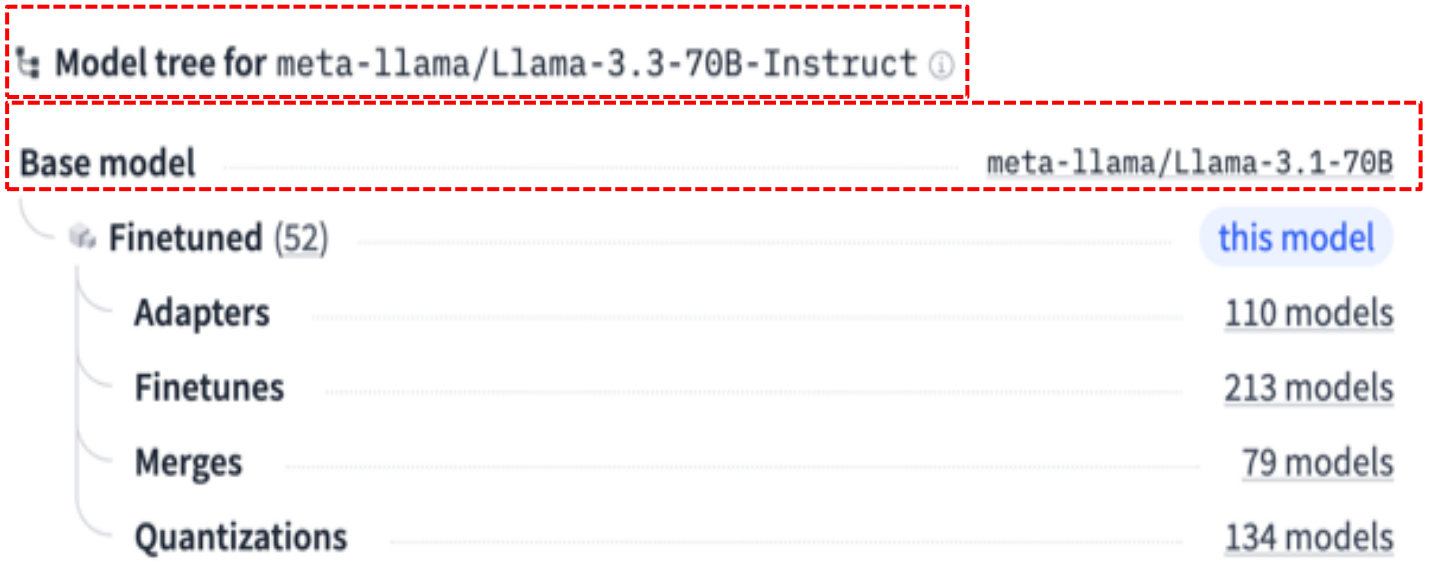
Overview of HuggingGraph

- HuggingGraph: a graph-based framework that analyzes the LLM supply chain between models and datasets.
  - *First work* to *systematically collect* the supply chain information of LLM.
  - *Construct a directed, heterogeneous graph* to analyze the relationship between models and datasets.
  - Performed *forward* and *backward* analysis to get the insights of the graph.

LLM Supply Chain Data Collection

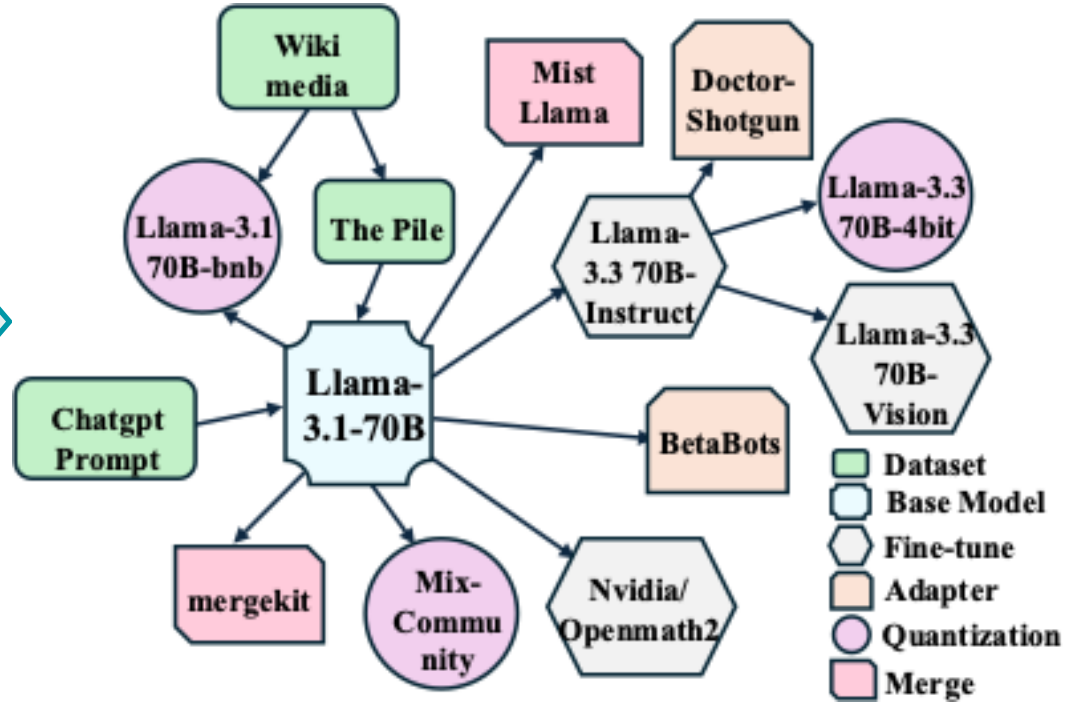
- Primarily use *four APIs* of Hugging Face for the LLM supply chain metadata collection.
  - Cross-link reference
  - Textual pattern recognition

API Name	Description
Model Hub	Access model hub to list, search, and download models and metadata
Datasets	Access the datasets for discovery, metadata retrieval, and downloading
Metrics	Access metrics for model evaluation
Search	Search name, tag, or other metadata for model and dataset



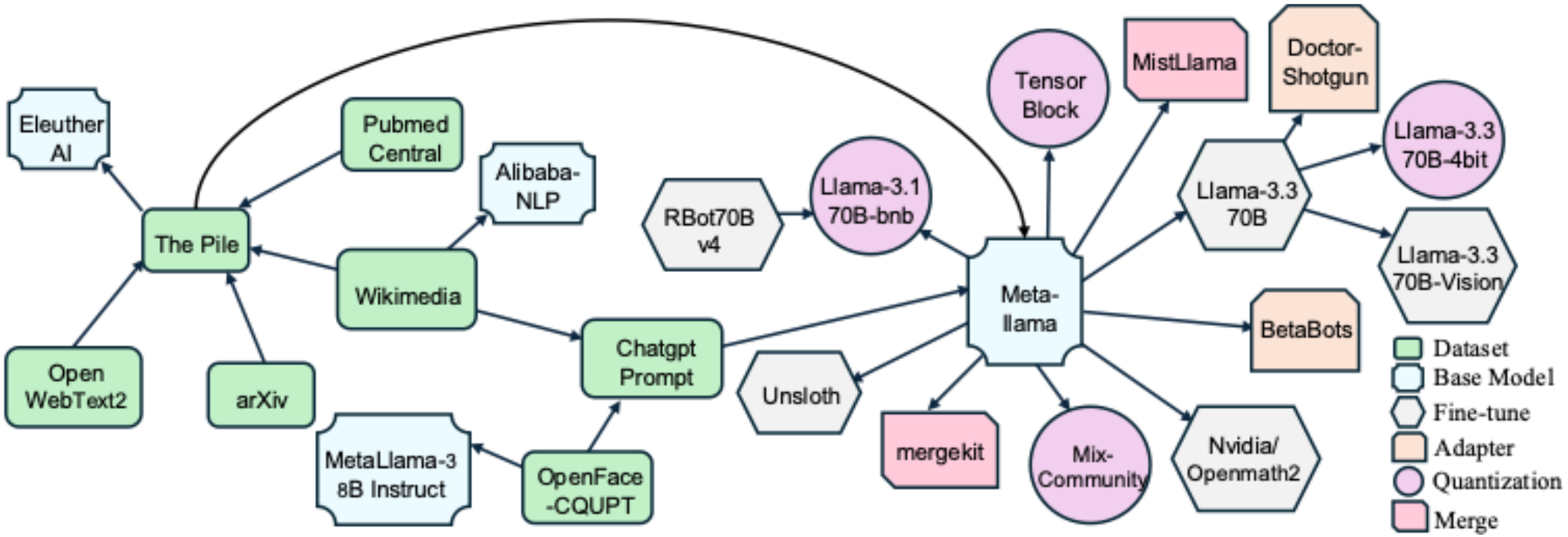
LLM Supply Chain Graph Construction

- Constructed a directed heterogeneous graph.
  - Model-model relationship
  - Dataset-dataset relationship
  - Model-dataset relationship



LLM Supply Chain Graph Analysis

- Analyzed model and dataset.
  - Forward Analysis: travers from a base model to its task-specific variant leaf model.
  - Backward Analysis: travers from a task-specific variant leaf model to its base model.



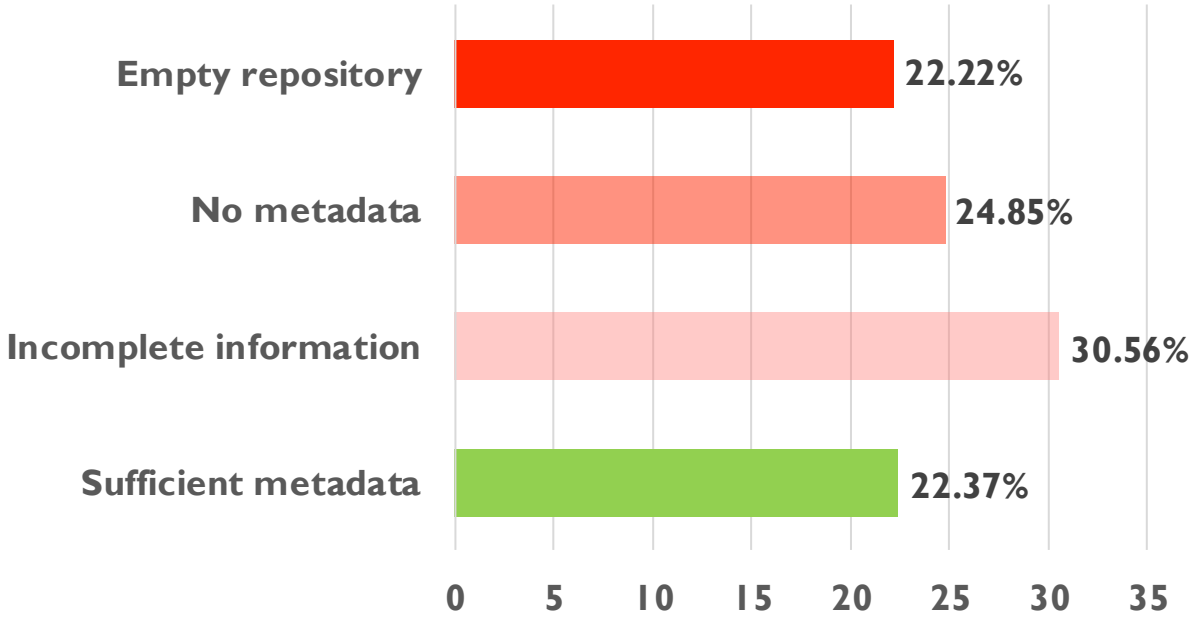
Experiment

- Experiment Setting
  - *NetworkX*: to construct and manage the directed heterogeneous graph.
  - *Algorithms*: Graph traversal, Connectivity, Community Detection.
- Research Questions:
  - RQ #1: What are the properties of LLM supply chain graph?
  - RQ #2: What structural patterns emerge?
  - RQ #3: What are the relationship between LLM models?
  - RQ #4: What are the relationships between models and datasets?
  - RQ #5: What insights can be gained from the dynamic updates?
  - RQ #6: How can HuggingGraph be applied to other platforms?

RQ #1: What are the properties of LLM supply chain graph?

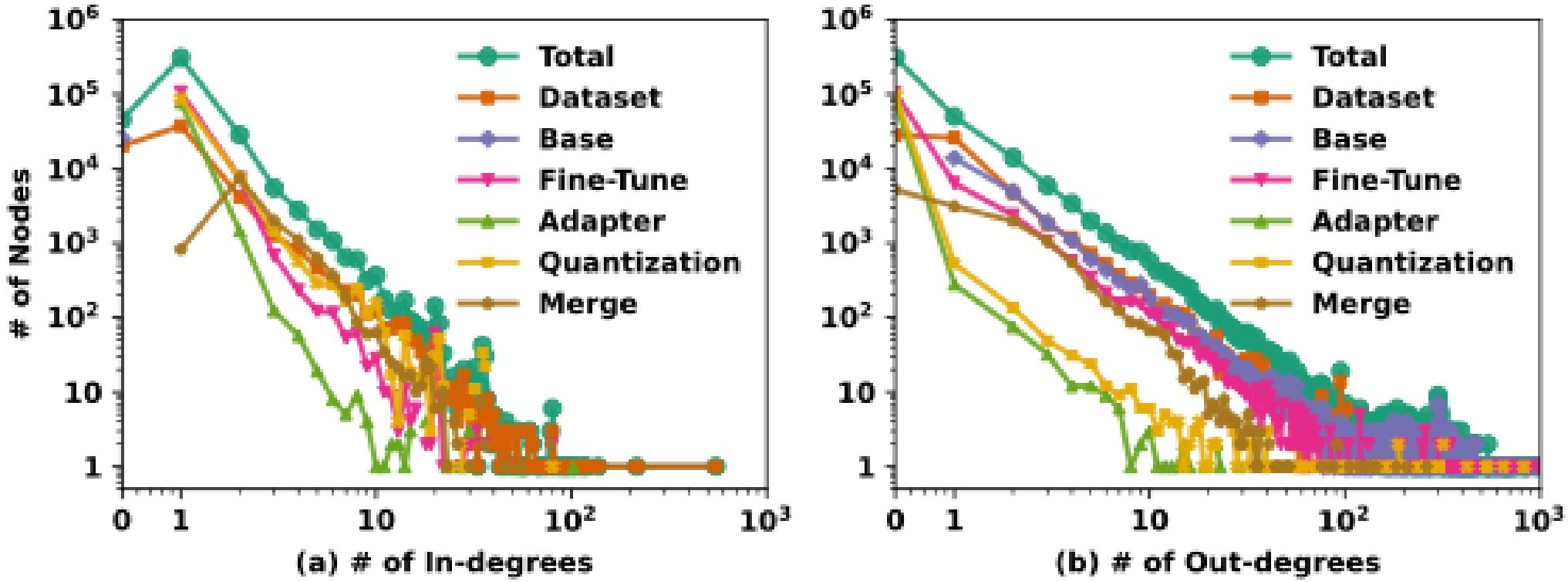
- The LLM supply chain graph is medium-scale and sparse.
  - Total # of nodes are 402,654, and edges 462,524 (as of June 30, 2025)
  - Six types of nodes: base, finetune, adapter, quantization, merge, and dataset.

Alarming findings about LLM supply chain metadata



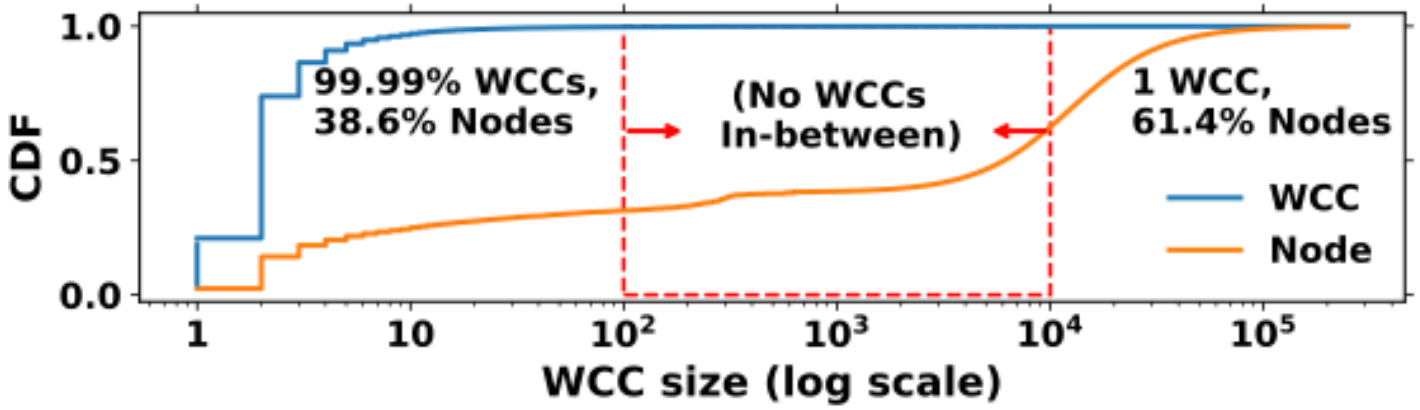
RQ #1: What are the properties of LLM supply chain graph?

- LLM supply chain graph – heavy-tailed distribution
- Most nodes have low degrees, and a few central nodes (hubs) dominate the graph



RQ #2: What structural patterns emerge?

- *Total #* of Weakly Connected Components (WCC) is 44,908, and the *largest* WCC covers 61.4% of all nodes.



- *Louvain Community Detection:* high modularity (0.96) indicating an extremely cohesive internal structure.

ID	Size	E.g. models	E.g. datasets	Modularity
1	9,390	OLMoE	Prompt-perfect	0.96
2	7,388	Qwen2.5_math	Marco-o1	0.96
3	6,989	Wanxiang	Smartllama3.1	0.96

- The LLM supply chain graph features a dominant core and semantically coherent communities.

RQ #3: What are the relationship between LLM models?

- Forward analysis to determine base model impact
  - Spawning thousands of derivative models.
- Downstream dependencies between the models within the LLM supply chain.

Base Model	Total	Fine-tune	Adapter	Quantization	Merge	Level
Llama-3.1-8B	7,544	1,710	1,542	3,473	1,693	25
Mistral-7B-v0.1	6,744	2,105	2,187	1,435	1,254	27
Qwen2.5-7B	6,733	1,972	1,764	2,516	1,132	11
Meta-Llama-3-8B	5,633	967	1,511	2,220	1,967	21
Llama-3.1-70B	4,063	698	281	2,075	2,519	11

RQ #3: What are the relationship between LLM models?

- Backward analysis to determine the task-specific model's impact
  - Exhibits deep dependencies with other task-specific variants.
- Upstream dependencies between the models within the LLM supply chain.

Model	Model Type	Total	Fine-tune	Level	Base model
Command-r-1-layer	Fine-tune	40	39	39	C4ai
KoModernBERT	Fine-tune	21	20	20	ModernBERT
T5-small	Fine-tune	21	20	20	T5-small
Clinical_260k	Fine-tune	20	19	19	Clinical_180K
T5-small-finetunned	Fine-tune	17	16	16	T5-small

RQ #4: What are the relationships between models & datasets?

- A single dataset is used to train multiple models.
- Datasets form the foundation of diverse model development in the LLM supply chain ecosystem.

Dataset	Total	Fine-tune	Adapter	Quantization	Merge
Mistral-v0.1	1,093	300	300	193	300
TinyLlama-1.1B-v1.0	728	300	300	100	28
Open_llama_3b	304	15	285	4	0
Yarn-Mistral-7b-128k	301	8	279	14	0
WizardVicuna-open-llama	280	12	261	7	0

RQ #4: What are the relationships between models & datasets?

- Multiple datasets are used to train one model.
- Models learn from multiple datasets to acquire knowledge, bridging diverse sources of information.

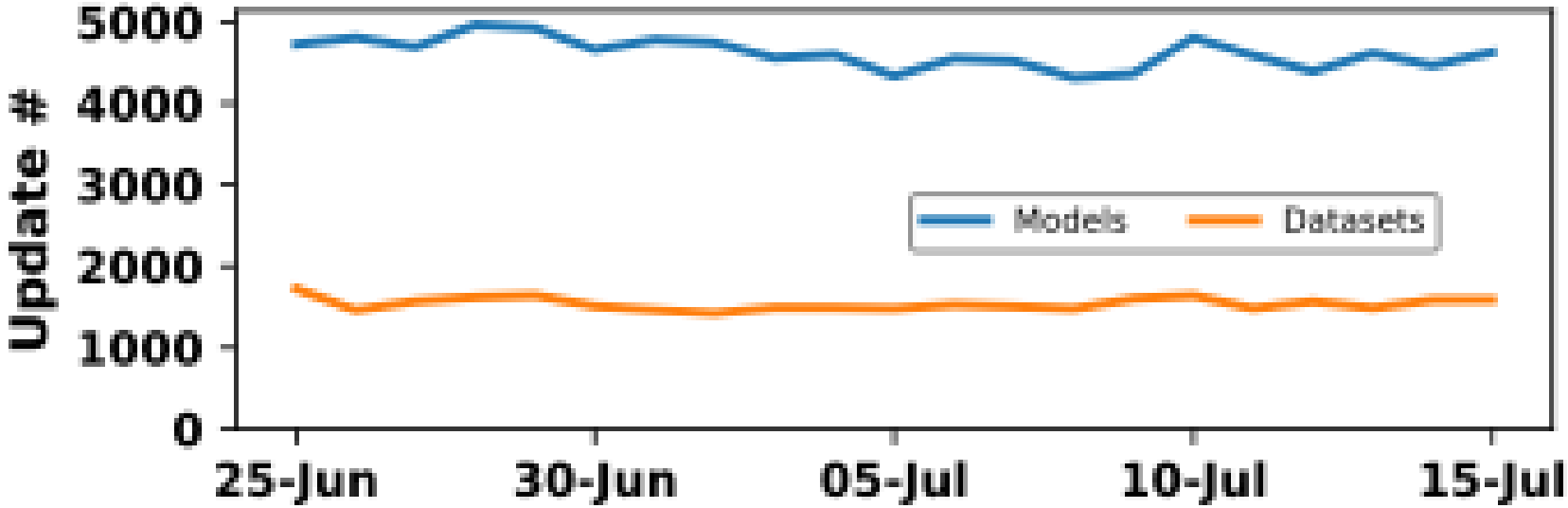
Model	Model Type	# of training dataset
DeBERTa-ST-AllLayers-v3.1	Finetune	116
DeBERTa-ST-AllLayers-v3.1bis	Adapter	116
Static-similarity-mrl-mul-v1	Finetune	108
Static-similarity-mrl-multilingual	Finetune	108
ModernBERT-base-embed	Finetune	88

RQ #5: What insights can be gained from the dynamic updates?

- Methodology:
  - Tracking and integrating new or deleted models and datasets on AI platforms to keep the LLM supply chain graph current.
    - Scope updates: Compare snapshots (t → t+1).
    - Collect metadata: Retrieve updated metadata.
    - Δ-Based Update: Refresh the graph as $G_{t+1} = G_t \cup \Delta_{t+1}$.

RQ #5: What insights can be gained from the dynamic updates?

- Average update per day (from June 25, 2025, to July 15, 2025)
  - # of models ~4,622, including *~3,843 additions* and *~779 deletions.*
  - # of datasets ~1,538, including *~1,350 additions* and *~233 deletions*.



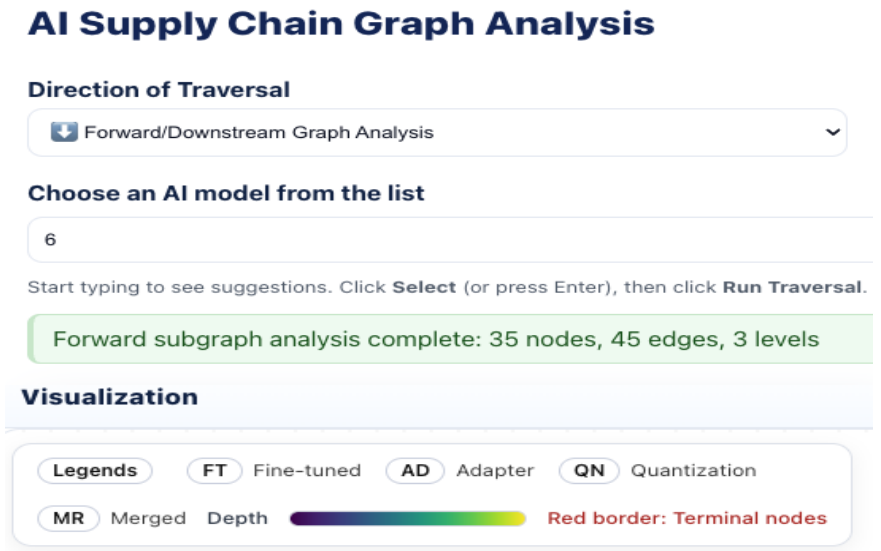
RQ #6: How can HuggingGraph be applied to other platforms?

- We applied our pipeline to another hosting platform, *Kaggle*
  - Total # of nodes are 2,777 and edges 3,990 (as of June 30, 2025)

- The resulting graph exhibits consistency with Hugging Face Analysis.
  - Heavy-tailed distribution
  - Sparse connectivity
  - Modular fragmentation

Demo Website for HuggingGraph

- We have developed the following website to visualize and analyze the LLM supply chain graph.
  - https://ai-supply-chain.github.io/



Key Takeaways

- *HuggingGraph,* the first LLM supply chain graph framework exploring the hidden LLM lineage.

- Deep understanding of how the LLM supply chain ecosystem evolves, interacts, and propagates risks.

Thank You!

- Graph Lab @ UTA
- **PhD positions available!**
- Webpage: https://yuede.github.io//lab
- GitHub repository: https://github.com/SC-Lab-Go/HuggingGraph