DataDetox: AC 215 Milestone 1

Kushal Chattopadhyay* Keyu Wang† Terry Zhou‡

October 14, 2025

1 Background and Motivation

While frontier and foundation models have achieved extraordinary results, the lack of transparency about their training data and upstream dependencies creates serious risks [3, 4]. Developers often inherit hidden flaws or biases from upstream models, and with increasing attention on AI safety and regulation, the urgency for practical lineage tracking is greater than ever. Bengio et al. even recommend that companies dedicate a third of their research budgets to safety-focused work [2].

We propose building an interactive, LLM-powered system that allows practitioners to query and explore model lineage through natural language Q&A. Instead of manually piecing together scattered information from model cards and papers, users can ask questions such as "Which models were trained on `LAION-5B` [11]?" or "I want to build a model that uses `Vicuna-13B` [13] and PLIP (Pathology-Language Pre-training) [7] from Nature — are there risks I should know about?" and receive clear, structured answers.

2 Problem Statement

The lack of transparency in how models are trained — both in terms of datasets and upstream dependencies — creates real risks for developers and organizations. Without accessible lineage information, it is difficult to identify harmful datasets, anticipate downstream impacts, or ensure compliance with emerging AI safety standards.

For example, the `LAION-5B` [11] dataset, widely used for training large vision–language models, has been shown to contain copyrighted images and sensitive personal data, raising legal and ethical concerns. Similarly, `MS-Celeb-1M` [6], once a popular face recognition dataset, was later withdrawn after researchers found it included images of private individuals without consent. Models trained on such datasets may silently propagate these risks downstream to countless derivative systems.

Manually piecing together this information from scattered model cards and research papers is inefficient and often incomplete. What is missing is a practical, interactive tool that makes lineage information easy to query and understand, so that practitioners can make informed decisions in real time.

3 Data

Source and Description

We will primarily rely on two data sources for this project: Hugging Face and arXiv. To trace the lineage of a given model, we will consult its model card on Hugging Face and extract information from the model tree (see Figure 1). This provides a clear view of both upstream and downstream models connected to the target model.

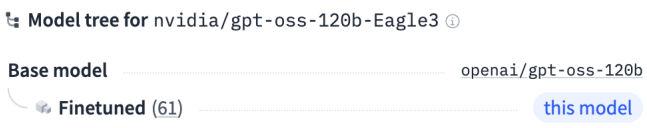

Figure 1: Model tree on Hugging Face for `nvidia/gpt-oss-120b-Eagle3`

*kchattopadhyay@college.harvard.edu
†keyuwang@g.harvard.edu
‡terryzhou@fas.harvard.edu

To gather information about a model's training data, we will examine its Hugging Face model card as well as any associated paper linked on the platform. We will then access the paper via arXiv and use a large language model (LLM) to parse it for relevant details concerning the model's training data.

These data are essential for creating a full dependency graph for each model.

Key Attributes

For model lineage data, each model may have upstream and downstream connections that collectively form a lineage tree. Internally, we will represent this tree in JSON format:

```
tree = {
    "model_A": {
        "children": {
            "model_B": {
                "children": {
                    "model_C": {"children": {}}
                }
            },
            "model_D": {"children": {}}
        }
    }
}
```

For model training data, we will parse the following attributes from Hugging Face model card or arXiv paper using LLM:

- `name` (`str`): Name of the dataset (e.g. Common Crawl)
- `human_synthetic` (`bool`): Whether the dataset is human generated or synthetic
- `synthetic_model` (`str`): If the data is synthetic, the model that generated the data
- `size` (`str`): Size of the dataset (e.g. 5B)

Data Quality

The primary data quality concern we anticipate is completeness. For many models — including some labeled as "open weight" — the training dataset is not publicly accessible. As a result, we may be unable to construct complete model dependencies. Nonetheless, we view the absence of training data itself as a meaningful signal worth capturing.

In addition, our model lineage information will be limited to what is available on Hugging Face. If certain upstream or downstream models are missing from the platform, we will not be able to recover them. However, given Hugging Face's prominence in the field, we do not expect this to be a major limitation.

Finally, since we will rely on large language models (LLMs) to parse training dataset details from Hugging Face model cards and associated arXiv papers, the accuracy of this information will depend heavily on the performance of the LLM. Optimizing for extraction quality will therefore be an important part of our project.

4 Scope and Objectives

We aim to build an interactive system that enables users to query and understand model lineages through natural language Q&A, powered by automated scraping and LLM-based parsing. Our objectives are:

- **Agentic scraping:** Develop a multi-step scraper that collects model trees and dataset information from Hugging Face and linked research papers.
- **LLM-based parsing:** Use large language models to extract and standardize lineage details from unstructured sources such as model cards and publications.
- **Interactive Q&A:** Implement a natural language interface where users can ask questions like "I want to build a model that uses Vicuna-13B and PLIP (Pathology-Language Pre-training) from Nature—are there risks I should know about?" This acts as an early-warning system, reducing the time engineers spend digging through scattered documentation before committing to a setup.
- **Visualization:** Provide clear, visual explanations of model dependencies, making lineage information accessible to both technical and non-technical users.
- **Real-world use cases:** Demonstrate how issues in datasets such as LAION-5B or MS-Celeb-1M propagate downstream into widely used models, highlighting the urgency and practical value of interactive lineage tracking.

5 Minimum Components for a Good Project.

- **Large or Heterogeneous Data:** Draws from diverse sources such as Hugging Face model cards and arXiv papers, requiring careful parsing and normalization.
- **Scalability:** Our backend will handle concurrent queries by caching common lineage trees and triggering on-demand scraping only when new models are requested. For queries to the LLM, we can also improve scalability by using asynchronous engineering structures, maintaining modularity in servers, and pre-training the model / establishing certain vector stores.
- **Complex Models:** Uses LLM agents to extract structured details from unstructured text, posing challenges in accuracy and multi-agent coordination.
- **Computationally Expensive Inference:** On-demand LLM parsing is resource-heavy, so efficiency techniques like batching, caching, and selective retrieval are essential.

6 Learning Emphasis.

The project emphasizes MCP usage to connect with complex pre-training model platforms like Hugging Face and RAG for external research information, as well as multi-agent architectures to provide specific risk mitigation recommendations, directly related to course concepts.

7 Application Mock Design.

We will utilize MCP development to connect with arXiv, Hugging Face, and other platforms that have access to relevant models and datasets. For development, we will ensure that our architecture reflects an optimized workflow using Multi-Agent Orchestration. We will use SFT (and other fine-tuning methods, like DPO) to improve response quality. Our wireframe is shown in Figure 2 and website mockup design for landing page and chatbot interface are shown in in Figure 3, 4, 5, and 6.

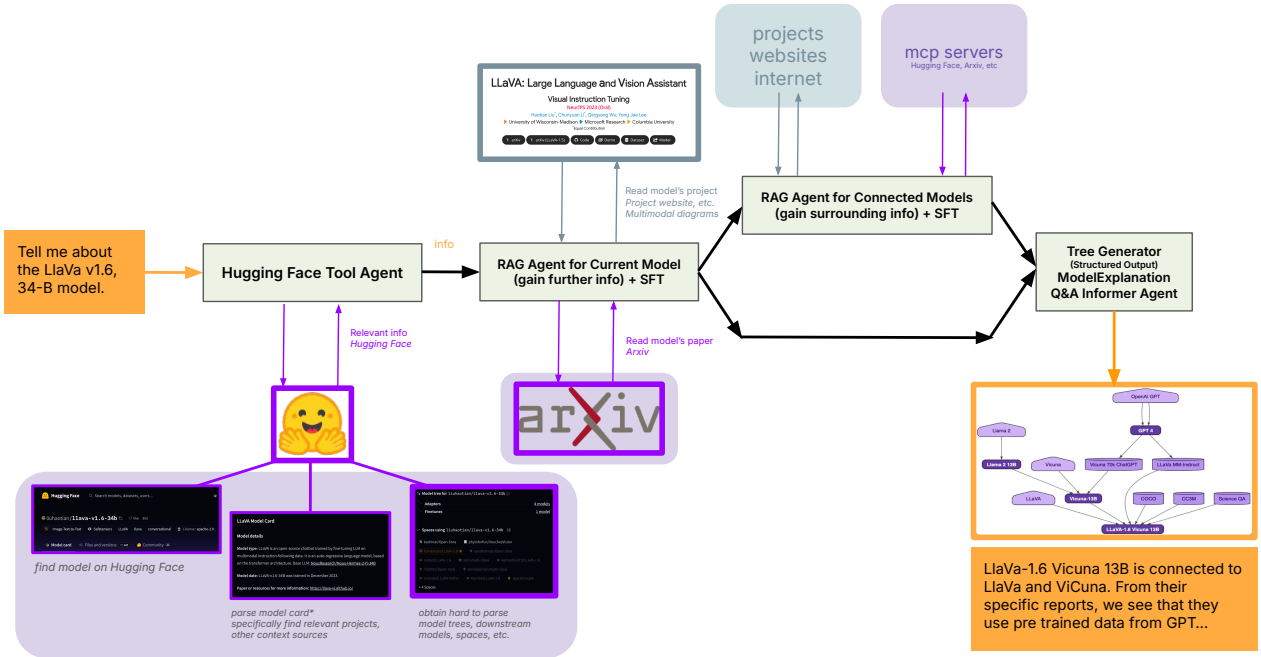

Figure 2: Overall architecture of DataDetox, Milestone 1

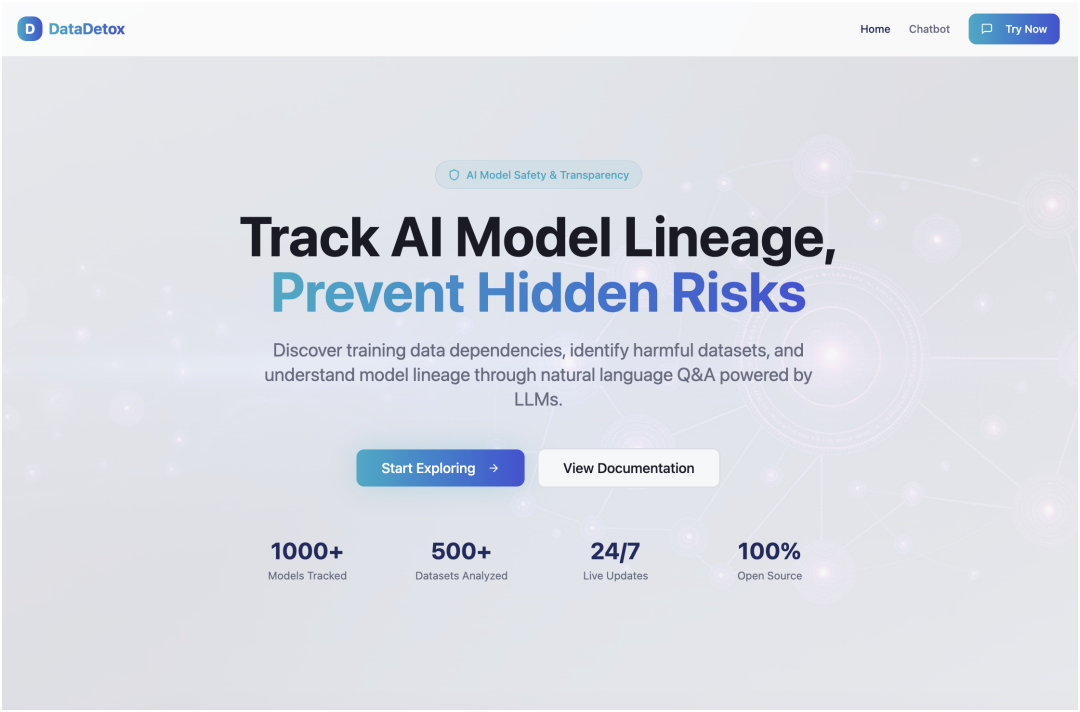

Figure 3: Top of Landing Page.

8 Research & Development

Prior work made efforts to improve transparency through model cards [10], datasheets [5], and metadata standards such as Croissant [1], but these do not fully capture upstream relationships. Hugging Face's "model trees" and ecosystem graphs [4] begin to address dependencies, while benchmarking platforms like HELM [9], HEIM [8], and ML Commons assess risks such as bias and safety, though they require heavy compute and cover only a limited set of models. Open platforms like Papers with Code and Hugging Face improve reproducibility but lack consistent standards. Most closely related to our work, Wang et al. propose model provenance as a way to mitigate downstream risks [12], but current efforts still do not provide an interactive, LLM-powered platform for exploring lineage in real time. Our system complements these initiatives by focusing on accessibility and interactivity, enabling practitioners to query and understand lineage directly.

9 Fun Factor

We are excited by the novelty of tracing model and data lineages, since it combines technical depth with practical relevance to AI safety and transparency. The interactive tree visualizations and the ability to "follow the data" across multiple models make the project engaging.

10 Limitations and Risks

Our approach faces several limitations, but we take steps to mitigate each:

- **Closed-source and API-only models:** Models like GPT-4 will remain out of scope since their training data and upstream dependencies are proprietary. We address this by focusing on open-source and open-weight models hosted on Hugging Face, where sufficient metadata is available to build meaningful lineage graphs.
- **Incomplete or vague documentation:** Mitigated by cross-parsing model cards, papers, and recording missing details explicitly as signals of risk in the database.

11 Milestones

We plan to structure the project into the following phases:

1. Agent development – earlier October

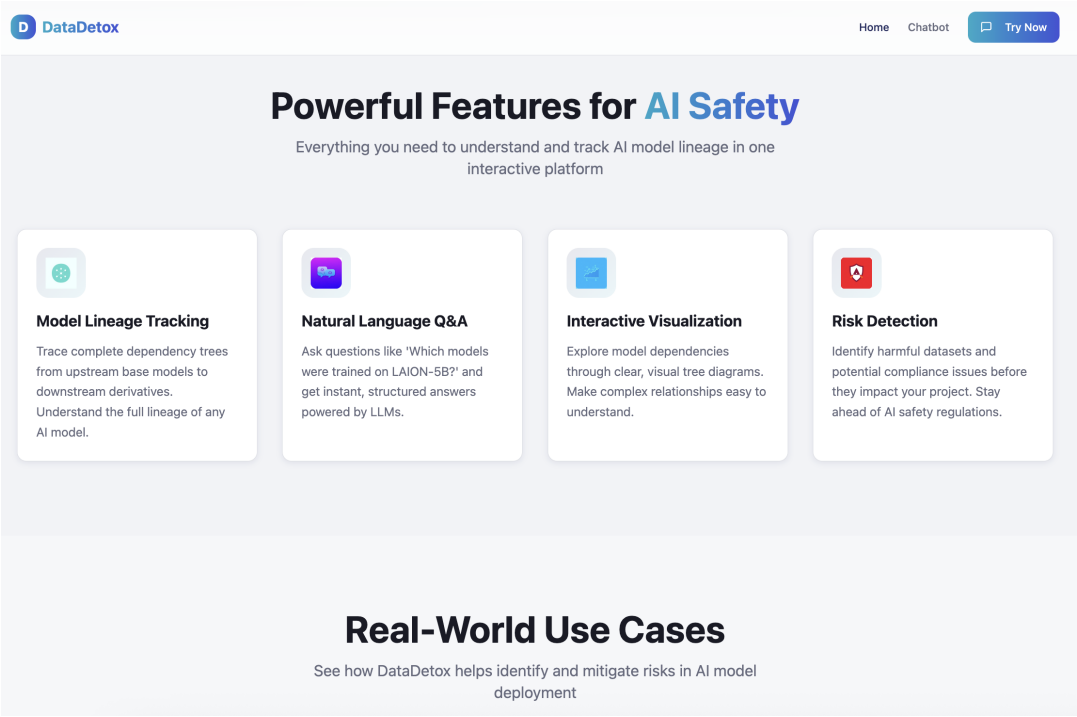

Figure 4: Middle of Landing Page.

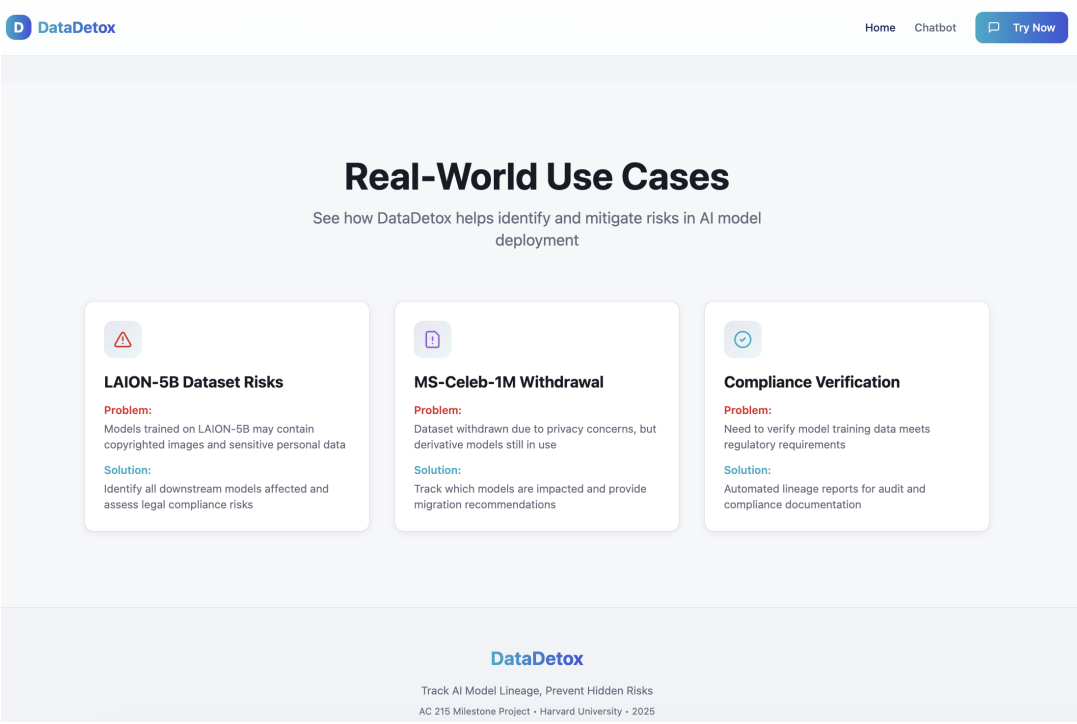

Figure 5: Bottom of Landing Page.

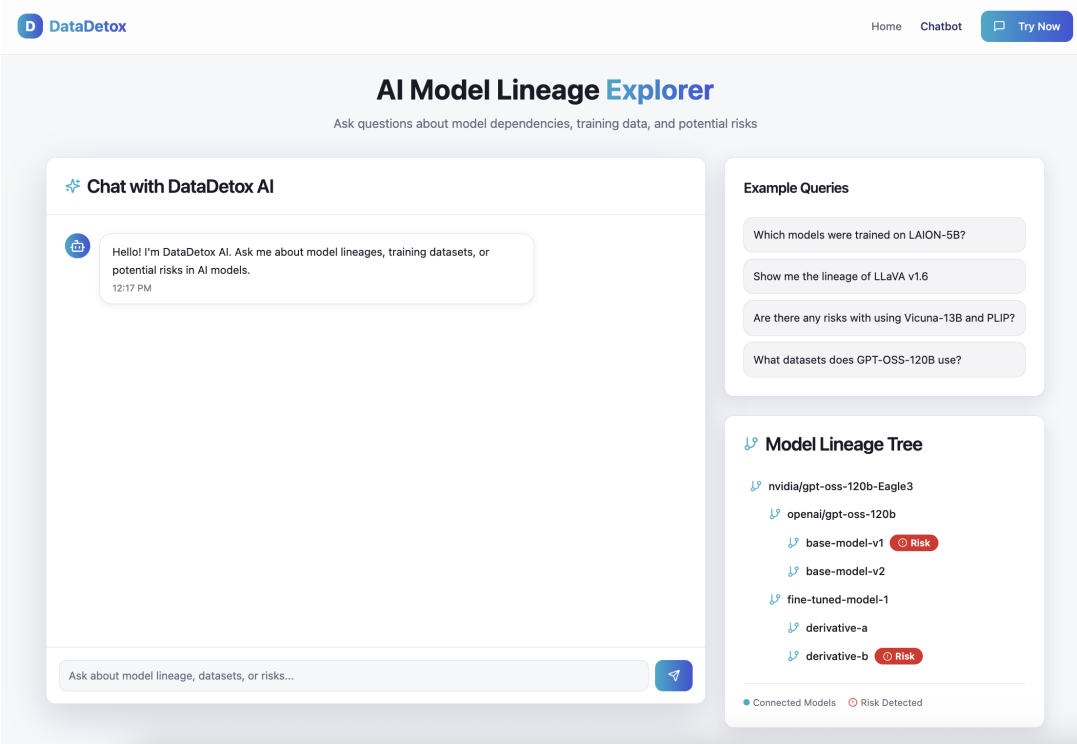

Figure 6: Chatbot interface.

2. Data collection and preprocessing – mid October
3. Backend implementation – early November
4. Frontend development – mid November
5. Chatbot integration – mid November
6. Final testing and deployment – late November, December

References

[1] Mubashara Akhtar et al. "Croissant: A metadata format for ml-ready datasets". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 82133–82148.

[2] Yoshua Bengio et al. "Managing extreme AI risks amid rapid progress". In: *Science* 384.6698 (2024), pp. 842–845.

[3] Rishi Bommasani. "On the opportunities and risks of foundation models". In: *arXiv preprint arXiv:2108.07258* (2021).

[4] Rishi Bommasani et al. "Ecosystem graphs: The social footprint of foundation models". In: *arXiv preprint arXiv:2303.15772* (2023).

[5] Timnit Gebru et al. "Datasheets for datasets". In: *Communications of the ACM* 64.12 (2021), pp. 86–92.

[6] Yandong Guo et al. "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition". In: *European conference on computer vision*. Springer. 2016, pp. 87–102.

[7] Zhi Huang et al. "A visual–language foundation model for pathology image analysis using medical Twitter". In: *Nature Medicine* 29 (2023), pp. 2307–2316. URL: https://www.nature.com/articles/s41591-023-02504-3.

[8] Tony Lee et al. "Holistic evaluation of text-to-image models". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 69981–70011.

[9] Percy Liang et al. "Holistic evaluation of language models". In: *arXiv preprint arXiv:2211.09110* (2022).

[10] Margaret Mitchell et al. "Model cards for model reporting". In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 220–229.

[11] Christoph Schuhmann et al. "Laion-5b: An open large-scale dataset for training next generation image-text models". In: *Advances in neural information processing systems* 35 (2022), pp. 25278–25294.

[12] Keyu Wang et al. "Mitigating Downstream Model Risks via Model Provenance". In: *arXiv preprint arXiv:2410.02230* (2024).

[13] Lianmin Zheng et al. "Judging llm-as-a-judge with mt-bench and chatbot arena". In: *Advances in neural information processing systems* 36 (2023), pp. 46595–46623.