Vision Paper

Model Lakes

Koyena Pal
Northeastern University
Boston, Massachusetts, USA
pal.k@northeastern.edu

David Bau
Northeastern University
Boston, Massachusetts, USA
davidbau@northeastern.edu

Renée J. Miller
Northeastern & U. Waterloo
Waterloo, ON, Canada
rjmiller@uwaterloo.ca

ABSTRACT

Given a set of deep learning models, it can be hard to find models appropriate to a task, understand the models, and characterize how models are different one from another. Currently, practitioners rely on manually-written documentation to understand and choose models. However, not all models have complete and reliable documentation. As the number of models increases, the challenges of finding, differentiating, and understanding models become increasingly crucial. Inspired from research on data lakes, we introduce the concept of *model lakes*. We formalize key model lake tasks, including *model attribution, versioning, search, and benchmarking*, and discuss fundamental research challenges in the management of large models. We also explore what data management techniques can be brought to bear on the study of large model management.

1 INTRODUCTION

With the dramatic rise in AI capabilities across a variety of domains [2, 3, 22, 26, 49, 50, 74, 94, 109, 125, 147], many organizations have begun to commit significant resources to developing Machine Learning Models. Many of these are fine-tuned versions of popular foundation models, such as Llama-3 [154], Mistral [63], DeepSeek-R1 [25], Stable Diffusion [121], BART [77], and BERT [27]. Proprietary closed-source models such as GPT-4 [110], Gemini [6] and Claude-3 [7] also support creation of fine-tuned models. To support this proliferation, sharing, and reuse of large models, many models are hosted on platforms to support the collaborative use and sharing of models such as Hugging Face [37] and Kaggle [66].

As the number of pre-trained models grows, comparing them and selecting the right one for specific tasks becomes increasingly difficult (see Example 1.1). Documentation, particularly model cards [97], aims to provide essential insights, but Liang et al. [80] have revealed a concerning lack of transparency and completeness in these resources. This makes it hard for users to make informed decisions, especially when navigating model sharing platforms. Efforts such as the Data Provenance Initiative [84], MLCommons [90, 118], and Responsible Foundation Model Development Cheatsheet [83] have been introduced to enhance the documentation of model creation and capabilities, with a particular focus on detailing their training data. However, not all model creators adhere to these guidelines, meaning that many existing and future models may still lack crucial information needed by users. To address this gap, we propose a systematic breakdown and formalization of tasks for "model lakes" — a system containing numerous heterogeneous pre-trained models and related data in their natural formats (one example is Hugging Face [37]).

© 2025 Copyright held by the owner/author(s). Published in Proceedings of the 28th International Conference on Extending Database Technology (EDBT), 25th March-28th March, 2025, ISBN 978-3-89318-099-8 on OpenProceedings.org.
Distribution of this paper is permitted under the terms of the Creative Commons license CC-by-nc-nd 4.0.

We introduce *model lakes*, as a parallel to data lakes, and discuss how important innovations in data lakes [106], including data discovery, annotation, and version management, can (and should) be applied within model lakes and studied with the same vigor. We look at how model lakes are currently managed and define and discuss tasks that can be used to better inform users about the models and their relationships. Others have discussed how one type of model (LLMs) may disrupt data management [41]. We focus on how data management can transform the management and use of AI models.

Example 1.1. **Model Search Problem:** Consider a situation where a user wants to find a model that can summarize a legal document and simplify it in a non-technical manner. On Hugging Face (as of Sept 2024), the user finds that there are around 1M+ models uploaded and 1950 of them have the 'summarization' task tag. While there are filters (trending, most likes, most downloads, model name search, and more), the user finds it hard to choose which model to use. There are various concerns that the user goes through as she scrolls through different model cards (a common semi-structured form of model documentation). Is this model aware of legal jargon? Is it good at summarizing and simplifying legal documents? Is this the latest version of the model? Was this model trained on legal texts and if so which texts? What are other models that are similar to this model? Are they also trained on the same or different legal texts?

2 THREE VIEWPOINTS OF A MODEL

An AI model, $\mathcal{M}$, can be analyzed from three viewpoints: according to its **history**, **intrinsic** composition, or **extrinsic** behavior. These viewpoints highlight different aspects of the model's characteristics, aspect that we will show are useful in analyzing model lake problems and solutions.

The **history** of the model is defined by its training data ($\mathcal{D}$) and training algorithm ($\mathcal{A}$), which may include processes like fine-tuning, model editing, or other adaptation techniques. The degree to which history is documented within a model lake can vary greatly [80].

The **intrinsic viewpoint** concerns the model's internal structure. This includes the model architecture (f_*), and the specific trained parameters (θ). The architecture refers to the structural design of the model, such as a combination of multi-layer perceptrons (MLPs) and attention layers in transformer models. Formally, the architecture defines a function family f_* which is instantiated with the specific learned parameters θ to create the function f_θ.

In contrast, the **extrinsic viewpoint** focuses on the model's observable behavior and performance in user-defined tasks. The model parameters, θ, are not visible as part of the extrinsics. However, the function f_θ and the model behavior, p_θ, is extrinsic. For example, in the case of an unconditional generative model, the extrinsics correspond to the observable probability distribution defined by the model, $p_\theta(x)$. For a classifier, the extrinsics are

defined by the behavior of the modeled predictions $p_\theta(y|x) = f_\theta(x)$. Extrinsics can be observed in terms of a neural network's action on inputs and outputs x and y, without requiring any knowledge of its training data or its internal structure.

The distinction among intrinsics, extrinsics, and history is useful because, while every model $\mathcal{M} = (\mathcal{D}, \mathcal{A}, f_*, \theta, p_\theta)$ has all these characteristics, there are cases where certain aspects may be unavailable. For example, in a Model Lake, the intrinsic details of some models might be inaccessible, and some analysis methods may rely solely on extrinsic observations or historical records to understand a model's behavior. We use this distinction to analyze possible solutions to a variety of model lake tasks. Our envisioned model lake is depicted in Figure 1.

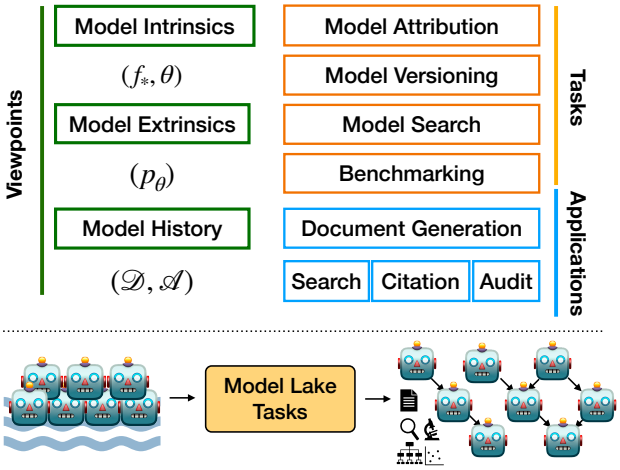

Figure 1: On the bottom of the figure, we illustrate the concept of *model lakes*, where diverse models are stored. As these models undergo the tasks outlined on the top-right side, users gain a deeper understanding of their origins, strengths, and how they are structured in relation to other models. This process provides key insights into the models' development, performance capabilities, and their positioning within the broader landscape of models. A model is defined as $\mathcal{M} = (\mathcal{D}, \mathcal{A}, f_*, \theta, p_\theta)$, where $\mathcal{D}$ (training data) and $\mathcal{A}$ (algorithm) can be traced through documentation, while architecture f_* and parameters θ come from accessible model weights, and behavior p_θ from observable outputs (illustrated on the upper left side of the figure).

3 FORMALIZING MODEL LAKE TASKS

Model lake tasks are specifically concerned with gathering and presenting insights about the models that are stored within the lake. Topics related to the infrastructure behind model training or storage systems fall outside this scope.

Model Attribution. In data lakes, data provenance is the "description of the origins of a piece of data and the process by which it arrived in a database" [17]. Model provenance (often called attribution) considers questions like "Why was image X generated when the model was given input Y?" If the history is recorded, the history can be consulted to directly ask the *training data attribution* question formally: which training data items $d \in \mathcal{D}$ are most influential on the decision; in other words, which d, if they were not present in the training data, would cause the decision to change the most? When history is not available, attribution can be studied using intrinsic and/or extrinsic clues to provide insight into the attribution of model decision behavior. For example, we can perform *sensitivity analysis* on the observable extrinsics of a model by asking: which aspects of the inputs to f_θ or p_θ are most important in a model's prediction of a particular output? And if we have access to model intrinsics, we can study *feature* or *representation analysis*, which asks, which internal representations or internal "concepts" within the model are most important for a decision?

Model Versioning. In the model versioning task, we want to understand whether (and possibly how) a model has been created from other models (similar to studying how a version of a data set or table may have been derived from another [136]). One possible definition of model versioning that uses an intrinsic viewpoint [56] is: Given a model, $\mathcal{M}_t$ and a set of N models, $\{\mathcal{M}_c | c \in N\}$, construct a directed Model Graph, $\mathcal{T}$, where a directed edge between models indicates that one model is a version of the other. The edges can describe the transformation. This can include training techniques, optimization techniques, as well as the data used to further train (fine-tune) models. An important problem in versioning is: given a model's training parameters θ_t and a candidate model's parameters, θ_c, is θ_c a source of θ_t? Being a source model would mean that θ_c or a version of θ_c was used for training θ_t.

Model Search. Model search is the task of finding a related or desired model. Again we can consider this task from different viewpoints. An extrinsic view considers the behavior of the model. Given a task function, $Q : X \rightarrow Y$ where a task takes an input, $x \in X$ and produces an output $y \in Y$, we want to find the best performing model, $\mathcal{M}_{best}$, for the given task among a set of N candidate models, $U = \mathcal{M}_1, \mathcal{M}_2, ...\mathcal{M}_N$. Each candidate model is characterized by their observable behavior, i.e., $\{p_{1\theta}, p_{2\theta}, ..., p_{N\theta}\}$. The optimal model is selected based on a scoring function of how close the model's behavior is to a query model.

Even considering only an extrinsic view, there are many formulations of model search. If the goal is to find models that perform similarly on specific data (for example, a specific image), then a possible definition is [85]: given a point, d, and a set of N candidate models, $U = \mathcal{M}_1, \mathcal{M}_2, ...\mathcal{M}_N$, rank models based on their similarity of their respective observable behavior $\{p_{1\theta}, p_{2\theta}, ..., p_{N\theta}\}$ w.r.t. to d. This could be extended to other types of models where for a given model, $\mathcal{M}_t$ and a set of candidate models, $\{\mathcal{M}_c | c \in n\}$, we want to find **related models** based on a ranking function that considers the similarity of models' output distribution and semantic concepts/patterns present within them.

An intrinsic view of model search would find models with similar model architectures and training parameters. Or as done in data lakes where intrinisic search is the norm, we could create embeddings representing the important features of the model and design a fast nearest neighbor search over these embeddings. Considering model history, we could search for models that have been trained on the same dataset. This is straight-forward when history is recorded, but when it is not fully explicit, we may leverage extrinsic or intrinsic clues in the search.

Benchmarking. For single model tasks, a benchmark, $\mathcal{B}$ (for example, a set of images), is used to measure the performance of a model, $\mathcal{M}$ (or set of models) based on a scoring function, $S(\mathcal{M}, \mathcal{B}) \in R$. For model lake tasks, we will need new (shared) model lake benchmarks (large sets of models $\mathcal{L}$) that can be used

to measure the quality of a lake task solution. This means that within a benchmark lake, we will need verified ground truth.

4 STATE-OF-THE-ART IN MODEL MGMT

Model Repositories, Registries, and Lakes. A model repository is a storage system for machine learning models. A model registry goes beyond basic storage by offering version control (the representation of versions, not today the discovery of version relationships mentioned in Section 4). These registries typically enforce clear naming conventions (for models and versions) and organize structured, standardized metadata [73, 149]. Recent advancements have added functionalities like explainable network intrusion detection [140] for security and integrations of different tools such as model monitoring and experiment tracking into a single framework [150]. Model registries are typically private, facilitating collaboration within organizations, while open platforms such as Civitai [61], Hugging Face [37], and Kaggle [66] enable public sharing of models. These platforms, which we can consider as model lakes, assist users in model exploration by providing curated catalogs and keyword search, often leveraging both manually created model names and metadata for more efficient discovery. Unlike lifecycle creation and training platforms, model lakes focus on managing a set of AI models, including their interrelationships. This is distinct from model management in databases, which deals with schemas and their mappings [13]. Current model lakes capabilities lack effective mechanisms for representing and navigating the model space semantically, particularly when model documentations or names are incomplete or unknown to users, leaving many models effectively undiscoverable [80].

Documenting Models. Models that are deployed are usually accompanied with documentation known as model cards [97]. Model cards contain (among other categories) information on model details, intended use, metrics, training data, and quantitative analyses. Similar to datasheets [45], model cards are designed to guide developers in documenting models in a structured way. Model cards can be and should be augmented with information more similar to nutritional labels [142] that also include information about fairness and bias in the data (models). They can be further enriched with lineage and security related documentation [48] such as adversarial attack and related defense measure policies, which are outlined in FactSheets for AI services [8].In addition to research on enhancing the completeness of model documentation, there remains a critical gap in the verification of model cards. There is a danger that people could intentionally misinform model users with malicious intent [130]. The state-of-the-art in verifying the documentation of a model is notably in its infancy [83]. Mithril AICert [129] is developing a certification initiative that verifies whether a model was trained using the specified algorithms and data sets. However, this initiative has some limitations. First, as it is still in development, it is not yet available for production use. Second, the certification process depends on the voluntary participation of the model creators. Furthermore, the AICert website [131] highlights several critical limitations, including the inability to audit training code or data for risks such as backdoors or data poisoning.

Model Search and Discovery. In Example 1.1, we present a potential scenario for the Model Search Problem. The current solution pipeline involves a user searching for a relevant model by naming specific models or by typing related keywords like `legal` to find models that either have that word in their name or in their model card. In other words, the search relies on the model's name and documentation. Hence, any sorting of the answer by relevance is just the relevance (prevalence) of the keywords and is not a semantic notion based on the model itself. Of course, this search may fail if the documentation is incomplete or inaccurate.

Within open data and enterprise data, researchers have learned that they cannot rely on metadata for datasets to be accurate, complete, or consistent within a data lake [106]. As a result, there is a great deal of work on semantic or content-based search in data lakes including join search [30, 68, 164, 165], union search [39, 59, 67, 107], and related dataset search [14, 40]. Machine learning models have revolutionized content-based dataset search. Important and impactful work has shown how we can use machine learning to create meaningful semantic representations of tuples, columns, and full tables to enhance dataset search and other semantic tasks like data integration and alignment.

But what about *content-based model search*? To the best of our knowledge, this is an area within ML that is in its infancy. Recent techniques for image models leverage meta learning [79, 85]. HuggingGPT [133] uses an LLM (in their case, ChatGPT) as a controller to decide which models to use based on a user's prompt. This metadata-based search differs from content-based approaches, as the LLM parses the user's prompt into tasks and uses model descriptions to select relevant models. While it allows queries across any modality or domain, it is limited by the LLM's capabilities and the quality of model documentation. Additionally, it may fail when object-centric tags do not fully capture the model. Our vision emphasizes that content-based model search must cover all models in model lakes, including large language models, while ensuring usability through speed and accuracy.

Another important problem is *related model search*. One approach to addressing related model search was explored by Lu et al. [85], who search for image generative models by using the behavior of another image model as a query. We propose extending such *model as query* searches to all models in a model lake to help users identify related models when exploring the model in question.

Attribution. An important line of inquiry considers questions related to provenance or attribution [91, 102]. These issues pose similar questions to those studied in the database community [18]: from where did a generated fact derive or why was a predication made? Building on the concept of data provenance [17], we can extend the notions of why-, where-, and how-provenance to models, under the name of *model attribution*. Similar research questions are also posed in model interpretability as part of local and global explanations. As in data provenance, the main issue is in whittling down all the provenance associated with some process (such as a full model) into simple, but useful explanations [18]. More than two decades of research in the data management community have produced elegant and simple, yet powerful, models for data provenance when the process is a query [15, 16] or workflow [98, 99]. In large model attribution, the goal cannot be to understand and track all inputs (data, hyperparameters, code, training regimes, ...) that were used to create a specific model output. Hence, the challenge is to find meaningful sets of concepts that can be tracked efficiently and that provide important insights into model behavior.

The training data attribution problem is nontrivial because every aspect of training has the potential to impact every decision of a model. A variety of methods have been developed to estimate

influence of training data on model behavior [52, 70, 153]. The current approaches require *extensive use of training data as well as costly analysis of model intrinsics*. Another approach to the problem is to apply techniques such as membership inference analyses, or membership inference attacks [134] that ask the question of whether a specific training data item d is present in the training data $\mathcal{D}$, or training data reconstruction methods that extract sets of items from the training data [19, 128].

Insight about the attribution of model behavior can also be studied through *model interpretability* methods: *extraction of relevant knowledge from a model concerning relationships either contained in data or learned by a model* [104]. Several works have surveyed techniques and research questions in this area [31, 47, 88, 103, 117, 138, 160]. These approaches can be broadly categorized into the following areas. Local model explanations explain the sensitivity of individual output predictions to local changes in inputs using gradients [143], masks [42, 114], local models [120], or Shapley values [21, 87]. Global explanations explain the mechanisms of the overall model at the level of attention patterns [23], representations [11, 166], circuits [36, 152], or neurons [9, 10, 46]. Models can also be designed to be inherently interpretable [122]. Lastly, datasets can be explained using natural language explanations [105], data visualizations [28], or by training inherently interpretable models [137, 148].

Model Versions. Models are valuable assets that can be adapted and reused to create new versions. The original or base version is typically a foundation model — a pre-trained model that learns general features from its training dataset, denoted as $\mathcal{D}$. By making further adjustments to the training algorithm ($\mathcal{A}$), architecture (f_*), or dataset ($\mathcal{D}$), subsequent versions of the base model can be developed. Common $\mathcal{A}$-based modifications include fine-tuning, parameter-efficient tuning, preference tuning, model stitching, and model editing. *Model stitching*, for example, involves altering f_* by combining the architectures of two or more models to create a hybrid model [76]. *Model editing* [44, 92, 93, 96, 111, 139] focuses on updating certain facts (e.g., changing the name of the current President of a country), making localized adjustments without retraining the entire model. *Fine-tuning* involves further training a model's parameters,θ, to improve performance on specific task(s) or domain(s). For instance, T5 [116] is a pre-trained model that has been trained on a collection of large collection of text (about 750 GB) to perform well on a diverse set of tasks. This has been further fine-tuned with the MIMIC-III [65] and IV [64] dataset to form Clinical-T5 [86] to perform better for medical domain-related tasks. In addition to the traditional fine-tuning strategy, *parameter-efficient fine-tuning* methods have emerged to reduce computational overhead by freezing most of the model's parameters, only updating a small subset necessary for fine-tuning. For instance, Low Rank Adaption (LoRA) [58] is a parameter-efficient fine-tuning method to adapt various models by only updating a low-rank subset [58]. *Preference tuning* is another advanced technique, as seen in ChatGPT [109] and InstructGPT [113], which integrates human feedback into the fine-tuning process [156]. Lastly, due to the emergent ability of models to perform tasks without training [29, 82], newer models leverage *prompting* as a way to control content generation without needing further updates to their parameters.

Information about model versioning can be inferred even if the model history is unavailable. For example, Mu et al. [102] propose a data- and model-driven method to encode "Model DNA" for identifying if a model is a pre-trained version of another, assuming both share the same architecture and training data. However, more challenging cases arise when the target model has a different architecture or is trained on only a subset of the source model's parameters. Hugging Face recently introduced new metadata fields in their model cards, enabling users to specify the base model and explain how it has been modified. This metadata generates a model tree, linking related models by their extensions. However, its accuracy depends on reliable documentation, and older models lack this data. Research on reconstructing relationship is emerging, such as Horwitz et al. [56]'s approach using weight similarities, though this approach is limited to known models with a single base version.

Privacy and Safety. Models are vulnerable to the disclosure of private information [19, 135] and adversarial results [78, 167] when attacked. Hence, initiatives such as Privacy Preserving Machine Learning [123] exist to understand, measure, and mitigate such risks. As a result, techniques such as differential privacy [33, 158], data sanitization [32, 141] and robust prompt optimization [162] have been utilized to defend against such attacks. These methods generally aim to obscure or eliminate private information while detecting and preventing attempts to jailbreak or manipulate the model and its output. However, this can create a false sense of privacy as defense techniques can continue to be compromised with other attack schemes [157].

Beyond privacy, ensuring AI systems are fundamentally safe is another critical challenge. Community-driven efforts focus on building safe AI [112], aiming to specify, verify, and ensure that the models behave as intended. A recent approach leverages neuroscience concepts, such as representation engineering, to enhance AI transparency [5, 95] and improve our understanding of traits like honesty, power seeking tendencies, and morality in models. However, this direction is still in its early stages, with open challenges including the scope of representations that can be explored and the development of effective evaluation methods to better inform users about model safety.

Benchmarking. Benchmarking plays a crucial role in evaluating model performance for specific tasks and remains a well-established research area, however benchmarks for newer topics such as model attribution and versioning are urgently needed. Developers frequently report model performance using standardized benchmarks, making it a routine part of model assessment. Broadly, there are two primary types of benchmarking: (1) evaluating how well a model, $\mathcal{M}$, approximates a ground truth distribution, and (2) assessing a model's performance against a targeted evaluation metric. In classification tasks, models are typically evaluated using accuracy, as ground-truth labels provide a direct assessment of the model's output. Additionally, confusion matrices offer insights into the types of errors made. For text generation tasks, perplexity is a widely used metric, and several popular benchmarks exist to assess performance [53, 54, 119]. In the case of image generation, the Fréchet Inception Distance (FID) [55] is a common metric and COCO [81] and VQAv2 [51] are some notable benchmarks in this domain. Beyond traditional performance metrics, benchmarking also considers biases related to protected attributes [126], as well as the environmental impact [75] of model training and deployment. Model lake benchmarks lack large-scale, publicly available datasets that mimic realistic conditions of diverse models in model lakes. There has been preliminary efforts by Lu et al. [85], where they released a benchmark for model search with 259 publicly available image

generative models and 1000 customized text-to-image diffusion models. Model Zoo [127] is another relatively large-scale dataset, but it is limited to vision models. Creating large-scale model lake benchmark that integrates various modalities (such as text, images, and audio) remains a research challenge.

5 RESEARCH ROADMAP

Building on current research and identified gaps outlined in Section 4, this section presents a vision to address these challenges. A model lake is illustrated in Figure 2, where rather than APIs, we envision data scientists interaction with the lake using queries.

Benchmarking. As discussed in Section 4, benchmarking, while a well-established research area, is under-explored in model lakes. An important topic is the development of lifelong benchmarks [115] that can address increasingly complex and novel scenarios as models continue to evolve in capability and diversity. In addition, there is a need to develop benchmarks specific to model lake tasks. For instance, to advance research in model attribution, a comprehensive *benchmark dataset* is needed—one that includes labeled model parameters, architectures, and detailed transformation records (e.g., fine-tuning, model editing). This benchmark can be extended to model versioning by adding data that tracks the previous and subsequent versions of models.

Model Inference. Deploying effective solutions for model lake tasks is a challenge with respect to *usability* and *scalability*. To improve usability, the model inference component involves *identifying appropriate benchmarks and generating relevant prompts, as well as selecting suitable models (target models or meta-models)* to address a user query. While users can manually run prompts and select models, this approach is prone to errors and suboptimal outcomes, especially if users lack the expertise to use models effectively. For example, a classifier's behavior may be misinterpreted if a user does not understand the type of data is was trained or the input it expects. To mitigate this, the proposed model lake tasks can incorporate additional perspectives, such as intrinsic model properties (e.g., weights), to provide insights and guide users towards a more accurate interpretations and applications. This search and generation process can also be automated using an AI agent. By applying benchmarks and model(s) in question, we combine research efforts of benchmarking and attribution questions, which are used to explain the behavior and output of a model.

Indexer. A central component of a model lake is the indexer, which would be used to embed and provide scalable sublinear search over the model embeddings. The indexer can use different ranking functions tailored to the task. Achille et al. [1], Wang et al. [154] have introduced an approach for generating model and task embeddings. However, search methods must scale to handle millions of models and adapt to newer models with advanced capabilities [155]. Indices like HNSW (Hierarchical New Small World) [89], have proven effective in practice in indexing high-dimensional embeddings enabling fast nearest-neighbor search (including over data lakes [39]. However, HNSW provides no formal guarantees on correctness and its use in model lakes remains under-explored. Effective embedding of models is crucial for accurate comparison and ranking by the indexer. Wang et al. [154] explores this but there work is not inclusive of all model types. A robust system should support diverse embeddings to ensure indexing effectiveness. Many of the model lake tasks will benefit from hybrid approach, that indexes both metadata and model embeddings – for example, related model search can combine well-chosen model embeddings representing important intrinsic model features with search over verified models cards. Similarly, in versioning, embeddings and their associated rankings can aid in identifying parent-child relationships and assessing the distinctiveness of each model relative to others.

Weight-Space Modeling. Weight-Space modeling is a hyper-representation learning approach where a neural network is trained to process weights of other models [34]. This method can be useful for making distinctions between models, especially in complex scenarios like model stitching, where similar models with multiple shared parent models need to be distinguished. This is a promising direction. Zhou et al. [163], for instance, reveals a linear connection between fine-tuned models. This approach could also facilitate dynamic selection of benchmarks for performance measurement by learning from previous relationships between datasets and models. The primary challenge in weight-space modeling is identifying the most relevant aspects of the model for training another model to uncover patterns, while simultaneously considering the weight-space model architecture [108]. It is also crucial to ensure sufficient data diversity to avoid overfitting or underfitting.

Interpretability. Interpretability methods can be most useful for the model attribution task where users must understand the origin of model behavior, for example, when detecting knowledge changes [159], or when unlearning learned knowledge [12, 35, 43, 62, 72]. Attribution questions also apply to the source model's architecture. Approaches like circuit discovery [24, 36, 152] can help identify the computational origin of model behavior. *Model inversion* can be used to recover an input prompt given an output [60, 100, 101]. These methods are also a promising route for understanding the impact of training data on internal model states. In recent work, Sharkey et al. [132] present a list of open questions in interpretability, many of which are relevant to the problem of model lakes.

Holistic Management of Models and Data. Effective model lakes need to integrate advances in data lakes in a holistic way given the reliance of models on data as their fuel. Many model management tasks include (in part) an analysis of training or input data. Hence, traditional data lake concerns such as data provenance, data versioning, and related issues still apply. Thus, in addition to new tasks specific to model lakes, we must also account for data lake tasks when dealing with the data used by or generated from these models. And integrating these tasks will be important. As a simple example, when searching for models trained on a dataset, users may want to find models trained on versions of the dataset.

6 MODEL LAKE TASKS: APPLICATIONS

Documentation Generation. To streamline the creation of detailed model cards, engineers can leverage model lake tasks to generate a rough draft of the required documentation. This application is similar to discovering and annotating metadata in data lakes [4, 38, 71, 124, 161]. Here's how the process might work: upon uploading a model to the model lake, state-of-the-art techniques for tasks like attribution, versioning, benchmarking, and others can automatically analyze and map the model's relationships to other models in the model lakes. The engineer can review and either accept or modify these generated mappings, especially if any information appears inaccurate, creating an initial version of the model details section. Additionally, the engineer can assess the model's robustness by testing its performance against

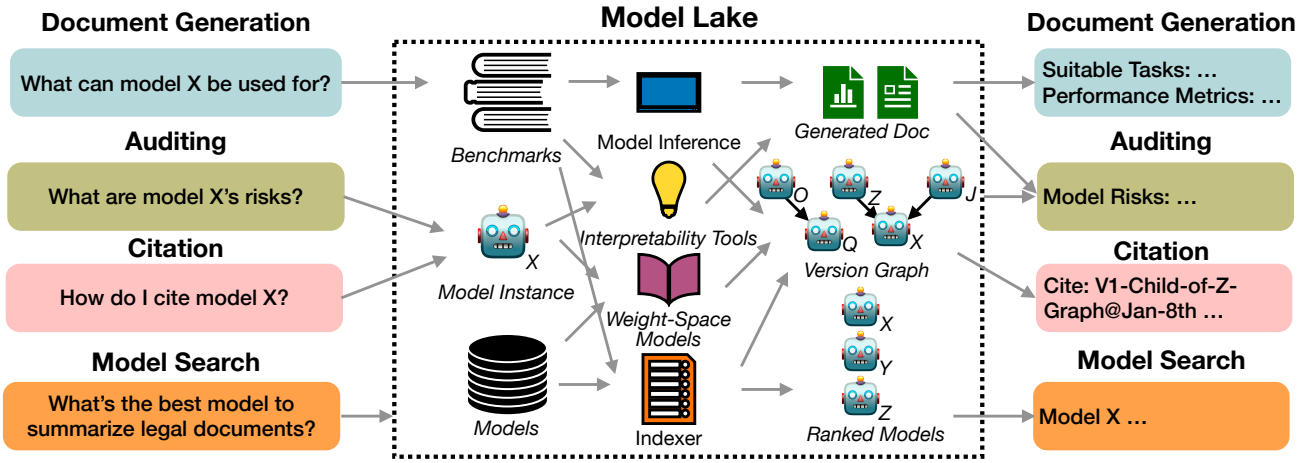

Figure 2: Model Lakes Design. A model lake stores models and processes them using techniques, like inference, interpretability, weight-space modeling and indexing to support various user interactions. It generates outputs like version graphs, model cards and ranked models, refining them into human-readable results, as shown on the figure's right side.

relevant benchmarks, including metrics like task accuracy, environmental impact, and more. Based on these test results, key sections of the model card, such as intended use and performance metrics, can be auto-populated for a faster, more reliable drafting process. It can alert developers to model risks, as shown by Wang et al. [151], who demonstrate how model versioning helps warn downstream model users of upstream model risks.

Auditing. Policymakers have recently proposed AI safety and compliance regulations [57, 144, 145] aimed at fostering more responsible and accountable AI models [20]. The model document generation application procedure can be repurposed for auditing by creating a template questionnaire [48] and using the information from the model lake to generate a draft response with proof or explanation about how a requirement is fulfilled. This process can incorporate privacy-related technical solutions, where insights from model lake tasks help identify vulnerabilities across related models and their successor versions. It can also aid in attributing sensitive data that the model may have access to, highlighting potential exposure risks.

Data and Model Citation. *Data citation* helps stakeholders identify the source, ownership, and authorship of the data used for a particular analysis. It is important to use proper data citation because the structure and contents of the database can evolve [16]. Hence, this problem is also extended to machine learning, since a large part of model training is its dataset. Thus, the task of citing data for data lakes remains relevant for documenting the training data used in the creation of the model. Similarly, *model citation* is essential, as users can further train the model or use its outputs for consumption or additional training. One proposed solution to identify generated output is the use of watermarks [69]. In addition, *model versioning* tasks provide crucial documentation, allowing researchers, engineers, and developers to refer to the exact version of the model used for training or content generation. If a particular model is used, the platform would refer to its versioning graph and generate a citation with the model version and timestamp of the graph. Upon any updates of the graph, a new citation would be generated with the updated version and timestamp. This would be useful for also citing any generated content from this model.

Model Search. In Example 1.1, we illustrate a potential scenario where a user seeks the most suitable model for summarizing and simplifying legal documents. As the model search task within the model lake evolves, we aim for users to be able to write declarative queries and retrieve a set of models ranked by their suitability for the specified task. Query example include "Find all models trained on this corpus of US Supreme Court cases" or "Find models that out perform Model X on Benchmark Y". Given a search task, the model lake framework can map the task function to a suitable indexer and run that indexer to retrieve top-ranked models. Whether the user is a technical or non-technical consumer, researcher, or engineer, they would be able to click on a model to view its model card. Attribution reveals how training data or learned concepts influence outputs, while version management clarifies the model's training process, lineage, and differences, enhancing transparency. Benchmarking evaluates the model's robustness on related tasks, addressing completeness. Together, these allow users to make informed decisions about the models they use.

7 CONCLUSION

The database community has responded to the "Big Model" revolution by proposing platforms like Agora [146], that manage data-related assets, including models, datasets, software, and compute resources in a coherent ecosystem. But the model-specific support in such an ecosystem must be expanded to include general methods for managing and finding models, support that can make such systems fully functional model lakes. It will be important that the methods generalize, irrespective of how many models we are trying to understand, what architectures the models use, how they are trained, or how we wish to query or search the models. We call on the database community to contribute to the vision of model lakes, supporting users to more easily find relevant models and to better understand those models. Our vision is for a fundamentally new platform that extends and integrates work on data/model attribution, data/model search, and data/model version management.

ACKNOWLEDGMENTS

KP and RM were supported in part by NSF award numbers IIS-2107248, IIS-1956096, IIS-2325632, and KP and DB by a grant from Open Philanthropy.

REFERENCES

[1] Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Stefano Soatto, and Pietro Perona. 2019. Task2Vec: Task Embedding for Meta-Learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 6429–6438. https://doi.org/10.1109/ICCV.2019.00653

[2] ELSA AI. [n.d.]. ELSA. https://elsaspeak.com/en/.

[3] Fitness AI. [n.d.]. Fitness AI. https://www.fitnessai.com/.

[4] Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Pieter Gijsbers, Joan Giner-Miguelez, Nitisha Jain, Michael Kuchnik, Quentin Lhoest, Pierre Marcenac, Manil Maskey, Peter Mattson, Luis Oala, Pierre Ruyssen, Rajat Shinde, Elena Simperl, Goeffry Thomas, Slava Tykhonov, Joaquin Vanschoren, Jos van der Velde, Steffen Vogler, and Carole-Jean Wu. 2024. Croissant: A Metadata Format for ML-Ready Datasets. In *Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning (DEEM '24)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3650203.3663326

[5] Sarah Chen James Campbell Phillip Guo Richard Ren Alexander Pan Xuwang Yin Mantas Mazeika Ann-Kathrin Dombrowski Shashwat Goel Nathaniel Li Michael J. Byun Zifan Wang Alex Mallen Steven Basart Sanmi Koyejo Dawn Song Matt Fredrikson Zico Kolter Dan Hendrycks Andy Zou, Long Phan. 2023. Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:cs.CL/2310.01405

[6] Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *CoRR* abs/2312.11805 (2023). https://doi.org/10.48550/ARXIV.2312.11805 arXiv:2312.11805

[7] Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf. Accessed: 2024-09-30.

[8] Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6–1.

[9] Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2018. Identifying and controlling important neurons in neural machine translation. *arXiv preprint arXiv:1811.01157* (2018).

[10] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6541–6549.

[11] Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics* 48, 1 (2022), 207–219.

[12] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. LEACE: Perfect linear concept erasure in closed form. arXiv:cs.LG/2306.03819

[13] Philip A. Bernstein and Sergey Melnik. 2007. Model management 2.0: manipulating richer mappings. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data (SIGMOD '07)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/1247480.1247482

[14] Alex Bogatu, Alvaro A. A. Fernandes, Norman W. Paton, and Nikolaos Konstantinou. 2020. Dataset Discovery in Data Lakes. In *ICDE*. 709–720.

[15] Peter Buneman, James Cheney, and Stijn Vansummeren. 2008. On the expressiveness of implicit provenance in query and update languages. *ACM Trans. Database Syst.* 33, 4, Article 28 (dec 2008), 47 pages. https://doi.org/10.1145/1412331.1412340

[16] Peter Buneman, Susan B. Davidson, and James Frew. 2016. Why data citation is a computational problem. *Commun. ACM* 59, 9 (2016), 50–57. https://doi.org/10.1145/2893181

[17] Peter Buneman, Sanjeev Khanna, and Tan Wang-Chiew. 2001. Why and where: A characterization of data provenance. In *Database Theory—ICDT 2001: 8th International Conference London, UK, January 4–6, 2001 Proceedings 8*. Springer, 316–330.

[18] Peter Buneman and Wang-Chiew Tan. 2018. Data Provenance: What next? *SIGMOD Rec.* 47, 3 (2018), 5–16. https://doi.org/10.1145/3316416.3316418

[19] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*. 2633–2650.

[20] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. 2024. Black-Box Access is Insufficient for Rigorous AI Audits. arXiv:cs.CY/2401.14446

[21] Hugh Chen, Ian C Covert, Scott M Lundberg, and Su-In Lee. 2023. Algorithms to estimate Shapley value feature attributions. *Nature Machine Intelligence* (2023), 1–12.

[22] Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware GPT-3 as a data generator for medical dialogue summarization. In *Machine Learning for Healthcare Conference*. PMLR, 354–372.

[23] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes (Eds.). Association for Computational Linguistics, Florence, Italy, 276–286. https://doi.org/10.18653/v1/W19-4828

[24] Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards Automated Circuit Discovery for Mechanistic Interpretability. arXiv:cs.LG/2304.14997

[25] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:cs.CL/2501.12948 https://arxiv.org/abs/2501.12948

[26] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2Web: Towards a Generalist Agent for the Web. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=kiYqbO3wqw

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[28] Victor Dibia. 2023. LIDA: A Tool for Automatic Generation of Grammar-Agnostic Visualizations and Infographics using Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Danushka Bollegala, Ruihong Huang, and Alan Ritter (Eds.). Association for Computational Linguistics, Toronto, Canada, 113–126. https://doi.org/10.18653/v1/2023.acl-demo.11

[29] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A Survey for In-context Learning. *arXiv preprint arXiv:2301.00234* (2022).

[30] Yuyang Dong, Chuan Xiao, Takuma Nozawa, Masafumi Enomoto, and Masafumi Oyamada. 2023. DeepJoin: Joinable Table Discovery with Pre-trained Language Models. *Proc. VLDB Endow.* 16, 10 (2023), 2458 – 2470. https://doi.org/10.14778/3603581.3603587

[31] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[32] Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. 2024. Reducing Privacy Risks in Online Self-Disclosures with Language Models. In *Proceedings of the 62nd Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 13732–13754. https://doi.org/10.18653/v1/2024.acl-long.741

[33] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer, 265–284.

[34] Gabriel Eilertsen, Daniel Jönsson, Timo Ropinski, Jonas Unger, and Anders Ynnerman. 2020. Classifying the classifier: dissecting the weight space of neural networks. In *ECAI 2020*. IOS Press, 1119–1126.

[35] Ronen Eldan and Mark Russinovich. 2023. Who's Harry Potter? Approximate Unlearning in LLMs. arXiv:cs.CL/2310.02238

[36] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread* (2021). https://transformer-circuits.pub/2021/framework/index.html.

[37] Hugging Face. 2023. https://huggingface.co

[38] Grace Fan, Jin Wang, Yuliang Li, and Renée J. Miller. 2023. Table Discovery in Data Lakes: State-of-the-art and Future Directions. In *Companion of the 2023 International Conference on Management of Data (SIGMOD '23)*. Association for Computing Machinery, New York, NY, USA, 69–75. https://doi.org/10.1145/3555041.3589409

[39] Grace Fan, Jin Wang, Yuliang Li, Dan Zhang, and Renée J. Miller. 2023. Semantics-aware Dataset Discovery from Data Lakes with Contextualized Column-based Representation Learning. *PVDLB* 16, 7 (2023), 1726–1739.

[40] Raul Castro Fernandez, Ziawasch Abedjan, Famien Koko, Gina Yuan, Samuel Madden, and Michael Stonebraker. 2018. Aurum: A Data Discovery System. In *34th IEEE International Conference on Data Engineering, ICDE 2018, Paris, France, April 16-19, 2018*. IEEE Computer Society, 1001–1012. https://doi.org/10.1109/ICDE.2018.00094

[41] Raul Castro Fernandez, Aaron J. Elmore, Michael J. Franklin, Sanjay Krishnan, and Chenhao Tan. 2023. How Large Language Models Will Disrupt Data Management. *Proc. VLDB Endow.* 16, 11 (jul 2023), 3302–3309. https://doi.org/10.14778/3611479.3611527

[42] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. 2019. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2950–2958.

[43] Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. 2023. Erasing Concepts from Diffusion Models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision*.

[44] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. 2024. Unified Concept Editing in Diffusion Models. *IEEE/CVF Winter Conference on Applications of Computer Vision* (2024).

[45] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[46] Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting Recall of Factual Associations in Auto-Regressive Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 12216–12235. https://aclanthology.org/2023.emnlp-main.751

[47] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael A. Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *5th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2018, Turin, Italy, October 1-3, 2018*, Francesco Bonchi, Foster J. Provost, Tina Eliassi-Rad, Wei Wang, Ciro Cattuto, and Rayid Ghani (Eds.). IEEE, 80–89. https://doi.org/10.1109/DSAA.2018.00018

[48] Delaram Golpayegani, Isabelle Hupont, Cecilia Panigutti, Harshvardhan J Pandit, Sven Schade, Declan O'Sullivan, and Dave Lewis. 2024. AI cards: towards an applied framework for machine-readable AI and risk documentation inspired by the EU AI Act. In *Annual Privacy Forum*. Springer, 48–72.

[49] Google. [n.d.]. BARD. https://bard.google.com/chat.

[50] Google. [n.d.]. Socratic. https://socratic.org/.

[51] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[52] Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. 2023. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296* (2023).

[53] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=d7KBjmI3GmQ

[54] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. https://openreview.net/forum?id=7Bywt2mQsCe

[55] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[56] Eliahu Horwitz, Asaf Shul, and Yedid Hoshen. 2024. On the Origin of Llamas: Model Tree Heritage Recovery. *arXiv preprint arXiv:2405.18432* (2024).

[57] The White House. 2022. *Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People*. The White House.

[58] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

[59] Xuming Hu, Shen Wang, Xiao Qin, Chuan Lei, Zhengyuan Shen, Christos Faloutsos, Asterios Katsifodimos, George Karypis, Lijie Wen, and Philip S. Yu. 2023. Automatic Table Union Search with Tabular Representation Learning. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*. Association for Computational Linguistics, 3786–3800. https://aclanthology.org/2023.findings-acl.233

[60] Xinting Huang, Madhur Panwar, Navin Goyal, and Michael Hahn. 2024. InversionView: A General-Purpose Method for Reading Information from Neural Activations. arXiv:cs.LG/2405.17653 https://arxiv.org/abs/2405.17653

[61] Civit AI Inc. 2023. https://civitai.com/

[62] Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge Unlearning for Mitigating Privacy Risks in Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 14389–14408. https://doi.org/10.18653/v1/2023.acl-long.805

[63] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *CoRR* abs/2310.06825 (2023). https://doi.org/10.48550/ARXIV.2310.06825 arXiv:2310.06825

[64] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data* 10, 1 (2023), 1.

[65] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.

[66] Kaggle. 2023. https://www.kaggle.com

[67] Aamod Khatiwada, Grace Fan, Roee Shraga, Zixuan Chen, Wolfgang Gatterbauer, Renée J. Miller, and Mirek Riedewald. 2023. SANTOS: Relationship-based Semantic Table Union Search. In *Accepted to appear in SIGMOD Conference*. ACM. https://arxiv.org/pdf/2209.13589.pdf

[68] Aamod Khatiwada, Roee Shraga, Wolfgang Gatterbauer, and Renée J Miller. 2022. Integrating Data Lake Tables. *Proceedings of the VLDB Endowment* 16, 4 (2022), 932–945.

[69] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A Watermark for Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.), Vol. 202. PMLR, 17061–17084. https://proceedings.mlr.press/v202/kirchenbauer23a.html

[70] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*. PMLR, 1885–1894.

[71] Keti Korini and Christian Bizer. 2023. Column Type Annotation using ChatGPT. In *Joint Proceedings of Workshops at the 49th International Conference on Very Large Data Bases (VLDB 2023), Vancouver, Canada, August 28 - September 1, 2023 (CEUR Workshop Proceedings)*, Rajesh Bordawekar, Cinzia Cappiello, Vasilis Efthymiou, Lisa Ehrlinger, Vijay Gadepally, Sainyam Galhotra, Sandra Geisler, Sven Groppe, Le Gruenwald, Alon Y. Halevy, Hazar Harmouch, Oktie Hassanzadeh, Ihab F. Ilyas, Ernesto Jiménez-Ruiz, Sanjay Krishnan, Tirthankar Lahiri, Guoliang Li, Jiaheng Lu, Wolfgang Mauerer, Umar Farooq Minhas, Felix Naumann, M. Tamer Özsu, El Kindi Rezig, Kavitha Srinivas, Michael Stonebraker, Satyanarayana R. Valluri, Maria-Esther Vidal, Haixun Wang, Jiannan Wang, Yingjun Wu, Xun Xue, Mohamed Zaït, and Kai Zeng (Eds.), Vol. 3462. CEUR-WS.org. https://ceur-ws.org/Vol-3462/TADA1.pdf

[72] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. 2023. Ablating Concepts in Text-to-Image Diffusion Models. In *International Conference on Computer Vision (ICCV)*.

[73] Neptune Labs. [n.d.]. neptune.ai | The MLOps stack component for experiment tracking. https://neptune.ai/.

[74] Prisma Labs. [n.d.]. Lensa. https://prisma-ai.com/lensa.

[75] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the Carbon Emissions of Machine Learning. *arXiv preprint arXiv:1910.09700* (2019).

[76] Karel Lenc and Andrea Vedaldi. 2015. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 991–999.

[77] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

[78] Tianlong Li, Zhenghua Wang, Wenhao Liu, Muling Wu, Shihan Dou, Changze Lv, Xiaohua Wang, Xiaoqing Zheng, and Xuanjing Huang. 2025. Revisiting Jailbreaking for Large Language Models: A Representation Engineering Perspective. In *Proceedings of the 31st International Conference on Computational Linguistics*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 3158–3178. https://aclanthology.org/2025.coling-main.212/

[79] Zhaotian Li, Binhang Qi, Hailong Sun, and Xiang Gao. 2023. AutoMRM: A Model Retrieval Method Based on Multimodal Query and Meta-learning *(CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 1228–1237. https://doi.org/10.1145/3583780.3614787

[80] Weixin Liang, Nazneen Rajani, Xinyu Yang, Ezinwanne Ozoani, Eric Wu, Yiqun Chen, Daniel Scott Smith, and James Zou. 2024. What's documented in AI? Systematic Analysis of 32K AI Model Cards. arXiv:cs.SE/2402.05160

[81] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[82] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.* 55, 9, Article 195 (Jan. 2023), 35 pages. https://doi.org/10.1145/3560815

[83] Shayne Longpre, Stella Biderman, Alon Albalak, Hailey Schoelkopf, Daniel McDuff, Sayash Kapoor, Kevin Klyman, Kyle Lo, Gabriel Ilharco, Nay San, Maribeth Rauh, Aviya Skowron, Bertie Vidgen, Laura Weidinger, Arvind Narayanan, Victor Sanh, David Ifeoluwa Adelani, Percy Liang, Rishi Bommasani, Peter Henderson, Sasha Luccioni, Yacine Jernite, and Luca Soldaini. 2024. The Responsible Foundation Model Development Cheatsheet: A Review of Tools & Resources. *CoRR* abs/2406.16746 (2024). https://doi.org/10.48550/ARXIV.2406.16746 arXiv:2406.16746

[84] Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi Wu, Enrico Shippole, Kurt D. Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, Deb Roy, and Sara Hooker. 2023. The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. *CoRR* abs/2310.16787 (2023). https://doi.org/10.48550/ARXIV.2310.16787 arXiv:2310.16787

[85] Daohan Lu, Sheng-Yu Wang, Nupur Kumari, Rohan Agarwal, Mia Tang, David Bau, and Jun-Yan Zhu. 2023. Content-based Search for Deep Generative Models. In *SIGGRAPH Asia 2023 Conference Papers (SA '23)*. ACM. https://doi.org/10.1145/3610548.3618189

[86] Qiuhao Lu, Dejing Dou, and Thien Nguyen. 2022. ClinicalT5: A Generative Language Model for Clinical Text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 5436–5443. https://doi.org/10.18653/v1/2022.findings-emnlp.398

[87] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *CoRR* abs/1705.07874 (2017). arXiv:1705.07874 http://arxiv.org/abs/1705.07874

[88] Haoyan Luo and Lucia Specia. 2024. From Understanding to Utilization: A Survey on Explainability for Large Language Models. *CoRR* abs/2401.12874 (2024). https://doi.org/10.48550/ARXIV.2401.12874 arXiv:2401.12874

[89] Yury A. Malkov and Dmitry A. Yashunin. 2020. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 4 (2020), 824–836.

[90] Peter Mattson, Christine Cheng, Cody Coleman, Greg Diamos, Paulius Micikevicius, David A. Patterson, Hanlin Tang, Gu-Yeon Wei, Peter Bailis, Victor Bittorf, David Brooks, Dehao Chen, Debojyoti Dutta, Udit Gupta, Kim M. Hazelwood, Andrew Hock, Xinyuan Huang, Bill Jia, Daniel Kang, David Kanter, Naveen Kumar, Jeffery Liao, Guokai Ma, Deepak Narayanan, Tayo Oguntebi, Gennady Pekhimenko, Lillian Pentecost, Vijay Janapa Reddi, Taylor Robie, Tom St. John, Carole-Jean Wu, Lingjie Xu, Cliff Young, and Matei Zaharia. 2019. MLPerf Training Benchmark. *CoRR* abs/1910.01500 (2019). arXiv:1910.01500 http://arxiv.org/abs/1910.01500

[91] Songzhu Mei, Cong Liu, Qinglin Wang, and Huayou Su. 2022. Model Provenance Management in MLOps Pipeline. In *Proceedings of the 2022 8th International Conference on Computing and Data Engineering (ICCDE '22)*. Association for Computing Machinery, New York, NY, USA, 45–50. https://doi.org/10.1145/3512850.3512861

[92] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and Editing Factual Associations in GPT. *Advances in Neural Information Processing Systems* 36 (2022).

[93] Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023. Mass Editing Memory in a Transformer. *The Eleventh International Conference on Learning Representations (ICLR)* (2023).

[94] Microsoft. [n.d.]. Seeing AI. https://www.microsoft.com/en-us/ai/seeing-ai.

[95] Patrick Mineault, Niccolò Zanichelli, Joanne Zichen Peng, Anton Arkhipov, Eli Bingham, Julian Jara-Ettinger, Emily Mackevicius, Adam Marblestone, Marcelo Mattar, Andrew Payne, Sophia Sanborn, Karen Schroeder, Zenna Tavares, and Andreas Tolias. 2024. NeuroAI for AI Safety. arXiv:cs.AI/2411.18526 https://arxiv.org/abs/2411.18526

[96] Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022. Fast Model Editing at Scale. In *International Conference on Learning Representations*. https://openreview.net/pdf?id=0DcZxeWfOPt

[97] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

[98] Luc Moreau and Paul Groth. 2013. *Provenance: An Introduction to PROV*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00528ED1V01Y201308WBE007

[99] Luc Moreau, Paolo Missier, Khalid Belhajjame, Reza B'Far, James Cheney, Sam Coppens, Stephen Cresswell, Yolanda Gil, Paul Groth, Graham Klyne, et al. 2013. Prov-dm: The prov data model.

[100] John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander Rush. 2023. Text Embeddings Reveal (Almost) As Much As Text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 12448–12460. https://doi.org/10.18653/v1/2023.emnlp-main.765

[101] John Xavier Morris, Wenting Zhao, Justin T Chiu, Vitaly Shmatikov, and Alexander M Rush. 2024. Language Model Inversion. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=t9dWHpGkPj

[102] Xin Mu, Yu Wang, Yehong Zhang, Jiaqi Zhang, Hui Wang, Yang Xiang, and Yue Yu. 2023. Model Provenance via Model DNA. arXiv:cs.LG/2308.02121

[103] Aaron Mueller, Jannik Brinkmann, Millicent L. Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. 2024. The Quest for the Right Mediator: A History, Survey, and Theoretical Grounding of Causal Interpretability. *CoRR* abs/2408.01416 (2024). https://doi.org/10.48550/ARXIV.2408.01416 arXiv:2408.01416

[104] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* 116, 44 (2019), 22071–22080. https://doi.org/10.1073/pnas.1900654116 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1900654116

[105] Avanika Narayan, Ines Chami, Laurel Orr, and Christopher Ré. 2022. Can Foundation Models Wrangle Your Data? *Proc. VLDB Endow.* 16, 4 (dec 2022), 738–746. https://doi.org/10.14778/3574245.3574258

[106] Fatemeh Nargesian, Erkang Zhu, Renée J. Miller, Ken Q. Pu, and Patricia C. Arocena. 2019. Data Lake Management: Challenges and Opportunities. *PVLDB* 12, 12 (2019), 1986–1989.

[107] Fatemeh Nargesian, Erkang Zhu, Ken Q Pu, and Renée J Miller. 2018. Table union search on open data. *PVLDB* 11, 7 (2018), 813–825.

[108] Aviv Navon, Aviv Shamsian, Idan Achituve, Ethan Fetaya, Gal Chechik, and Haggai Maron. 2023. Equivariant Architectures for Learning in Deep Weight Spaces. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.), Vol. 202. PMLR, 25790–25816. https://proceedings.mlr.press/v202/navon23a.html

[109] OpenAI. [n.d.]. Chat GPT. https://chat.openai.com/chat.

[110] OpenAI. 2023. GPT-4 Technical Report. *CoRR* abs/2303.08774 (2023). https://doi.org/10.48550/ARXIV.2303.08774 arXiv:2303.08774

[111] Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. 2023. Editing Implicit Assumptions in Text-to-Image Diffusion Models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 7030–7038. https://doi.org/10.1109/ICCV51070.2023.00649

[112] Pedro A Ortega, Vishal Maini, and DeepMind Safety Team. 2018. Building safe artificial intelligence: specification, robustness, and assurance. *DeepMind Safety Research Blog* (2018).

[113] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:cs.CL/2203.02155 https://arxiv.org/abs/2203.02155

[114] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*

(2018).
[115] Ameya Prabhu, Vishaal Udandarao, Philip Torr, Matthias Bethge, Adel Bibi, and Samuel Albanie. 2024. Lifelong Benchmarks: Efficient Model Evaluation in an Era of Rapid Progress. arXiv:cs.LG/2402.19472 https://arxiv.org/abs/2402.19472
[116] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.
[117] Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. 2022. Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks. *CoRR* abs/2207.13243 (2022). https://doi.org/10.48550/ARXIV.2207.13243 arXiv:2207.13243
[118] Vijay Janapa Reddi, Christine Cheng, David Kanter, Peter Mattson, Guenther Schmuelling, Carole-Jean Wu, Brian Anderson, Maximilien Breughe, Mark Charlebois, William Chou, Ramesh Chukka, Cody Coleman, Sam Davis, Pan Deng, Greg Diamos, Jared Duke, Dave Fick, J. Scott Gardner, Itay Hubara, Sachin Idgunji, Thomas B. Jablin, Jeff Jiao, Tom St. John, Pankaj Kanwar, David Lee, Jeffery Liao, Anton Lokhmotov, Francisco Massa, Peng Meng, Paulius Micikevicius, Colin Osborne, Gennady Pekhimenko, Arun Tejusve Raghunath Rajan, Dilip Sequeira, Ashish Sirasao, Fei Sun, Hanlin Tang, Michael Thomson, Frank Wei, Ephrem Wu, Lingjie Xu, Koichi Yamada, Bing Yu, George Yuan, Aaron Zhong, Peizhao Zhang, and Yuchen Zhou. 2019. MLPerf Inference Benchmark. *CoRR* abs/1911.02549 (2019). arXiv:1911.02549 http://arxiv.org/abs/1911.02549
[119] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. In *First Conference on Language Modeling*. https://openreview.net/forum?id=Ti67584b98
[120] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
[121] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042
[122] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. 2022. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys* 16 (2022), 1–85.
[123] Victor Ruehle, Robert Sim, Sergey Yekhanin, Nishanth Chandran, Melissa Chase, Daniel Jones, Kim Laine, Boris Köpf, Jaime Teevan, Jim Kleewein, et al. 2021. Privacy preserving machine learning: Maintaining confidentiality and preserving trust. https://www.microsoft.com/en-us/research/blog/privacy-preserving-machine-learning-maintaining-confidentiality-and-preserving-trust/
[124] Pegdwendé Sawadogo and Jérôme Darmont. 2021. On data lake architectures and metadata management. *Journal of Intelligent Information Systems* 56 (2021), 97–120.
[125] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2024. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems* 36 (2024).
[126] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22522–22531.
[127] Konstantin Schürholt, Diyar Taskiran, Boris Knyazev, Xavier Giró-i Nieto, and Damian Borth. 2022. Model zoos: A dataset of diverse populations of neural network models. *Advances in Neural Information Processing Systems* 35 (2022), 38134–38148.
[128] Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary C. Lipton, and J. Zico Kolter. 2024. Rethinking LLM Memorization through the Lens of Adversarial Compression. arXiv:cs.LG/2404.15146 https://arxiv.org/abs/2404.15146
[129] Mithril Security. [n.d.]. AICert – Open-source tool to trace AI models' provenance. https://www.mithrilsecurity.io/aicert.
[130] Mithril Security. [n.d.]. PoisonGPT: How We Hid a Lobotomized LLM on Hugging Face to Spread Fake News. https://blog.mithrilsecurity.io/poisongpt-how-we-hid-a-lobotomized-llm-on-hugging-face-to-spread-fake-news/.
[131] Mithril Security. [n.d.]. Welcome to AICert! https://aicert.mithrilsecurity.io/en/latest/.
[132] Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. 2025. Open Problems in Mechanistic Interpretability. arXiv:cs.LG/2501.16496 https://arxiv.org/abs/2501.16496
[133] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2024. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems* 36 (2024).
[134] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting Pretraining Data from Large Language Models. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=zWqr3MQuNs
[135] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks Against Machine Learning Models. In *2017 IEEE Symposium on Security and Privacy (SP)*. 3–18. https://doi.org/10.1109/SP.2017.41
[136] Roee Shraga and Renée J. Miller. 2023. Explaining Dataset Changes for Semantic Data Versioning with Explain-Da-V. *Proc. VLDB Endow.* 16, 6 (2023), 1587–1600. https://doi.org/10.14778/3583140.3583169
[137] Chandan Singh, Armin Askari, Rich Caruana, and Jianfeng Gao. 2023. Augmenting interpretable models with large language models during training. *Nature Communications* 14, 1 (2023), 7913.
[138] Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. Rethinking Interpretability in the Era of Large Language Models. arXiv:cs.CL/2402.01761
[139] Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitry Pyrkin, Sergei Popov, and Artem Babenko. 2020. Editable Neural Networks. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HJedXaEtvS
[140] Vincenzo Spadari, Francesco Cerasuolo, Giampaolo Bovenzi, and Antonio Pescapè. 2024. An MLOps Framework for Explainable Network Intrusion Detection with MLflow. In *2024 IEEE Symposium on Computers and Communications (ISCC)*. 1–6. https://doi.org/10.1109/ISCC61673.2024.10733700
[141] Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. 2025. Language Models are Advanced Anonymizers. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=82p8VHRsaK
[142] Julia Stoyanovich and Bill Howe. 2019. Nutritional Labels for Data and Models. *IEEE Data Eng. Bull.* 42, 3 (2019), 13–23. http://sites.computer.org/debull/A19sept/p13.pdf
[143] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.
[144] Elham Tabassi. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). https://doi.org/10.6028/NIST.AI.100-1
[145] Elham Tabassi. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance) PE/24/2024/REV/1, OJ L, 2024/1689, 12.7.2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj
[146] Jonas Traub, Zoi Kaoudi, Jorge-Arnulfo Quiané-Ruiz, and Volker Markl. 2020. Agora: Bringing Together Datasets, Algorithms, Models and More in a Unified Ecosystem [Vision]. *SIGMOD Rec.* 49, 4 (2020), 6–11. https://doi.org/10.1145/3456859.3456861
[147] Immanuel Trummer. 2022. CodexDB: Synthesizing code for query processing from natural language instructions using GPT-3 Codex. *Proceedings of the VLDB Endowment* 15, 11 (2022), 2921–2928.
[148] Berk Ustun and Cynthia Rudin. 2016. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning* 102 (2016), 349–391.
[149] Sandeep Verma, Farooq Sabir, Mani Khanuja, Rupinder Grewal, Saumitra Vikram, and Sreedevi Srinivasan. 2022. Build a cross-account MLOps workflow using the Amazon SageMaker model registry. https://aws.amazon.com/blogs/machine-learning/build-a-cross-account-mlops-workflow-using-the-amazon-sagemaker-model-registry/.
[150] T Vishwambari and Sonali Agrawal. 2023. Integration of Open-Source Machine Learning Operations Tools into a Single Framework. In *2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*. 335–340. https://doi.org/10.1109/ICCCIS60361.2023.10425558
[151] Keyu Wang, Abdullah Norozi Iranzad, Scott Schaffter, Doina Precup, and Jonathan Lebensold. 2024. Mitigating Downstream Model Risks via Model Provenance. arXiv:cs.LG/2410.02230 https://arxiv.org/abs/2410.02230
[152] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small. In *The Eleventh International Conference on Learning Representations*.
[153] Sheng-Yu Wang, Alexei A Efros, Jun-Yan Zhu, and Richard Zhang. 2023. Evaluating data attribution for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7192–7203.
[154] Wenxiao Wang, Weiming Zhuang, and Lingjuan Lyu. 2024. Towards Fundamentally Scalable Model Selection: Asymptotically Fast Update and Selection. *CoRR* abs/2406.07536 (2024). https://doi.org/10.48550/ARXIV.2406.07536 arXiv:2406.07536
[155] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research* (2022). https://openreview.net/forum?id=yzkSU5zdwD Survey Certification.
[156] Genta Indra Winata, Hanyang Zhao, Anirban Das, Wenpin Tang, David D. Yao, Shi-Xiong Zhang, and Sambit Sahu. 2024. Preference Tuning with Human Feedback on Language, Speech, and Vision Tasks: A Survey.

arXiv:cs.CL/2409.11564 https://arxiv.org/abs/2409.11564

[157] Rui Xin, Niloofar Mireshghallah, Shuyue Stella Li, Michael Duan, Hyunwoo Kim, Yejin Choi, Yulia Tsvetkov, Sewoong Oh, and Pang Wei Koh. 2024. A False Sense of Privacy: Evaluating Textual Data Sanitization Beyond Surface-level Privacy Leakage. In *Neurips Safe Generative AI Workshop 2024*. https://openreview.net/forum?id=3JLtuCozOU

[158] Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. 2024. On protecting the data privacy of large language models (llms): A survey. *arXiv preprint arXiv:2403.05156* (2024).

[159] Paul Youssef, Zhixue Zhao, Jörg Schlötterer, and Christin Seifert. 2024. Detecting Edited Knowledge in Language Models. arXiv:cs.CL/2405.02765 https://arxiv.org/abs/2405.02765

[160] Haiyan Zhao, Fan Yang, Himabindu Lakkaraju, and Mengnan Du. 2024. Opening the Black Box of Large Language Models: Two Views on Holistic Interpretability. *CoRR* abs/2402.10688 (2024). https://doi.org/10.48550/ARXIV.2402.10688 arXiv:2402.10688

[161] Yan Zhao, Franck Ravat, Julien Aligon, Chantal Soule-dupuy, Gabriel Ferrettini, and Imen Megdiche. 2021. Analysis-oriented Metadata for Data Lakes. In *Proceedings of the 25th International Database Engineering & Applications Symposium (IDEAS '21)*. Association for Computing Machinery, New York, NY, USA, 194–203. https://doi.org/10.1145/3472163.3472273

[162] Andy Zhou, Bo Li, and Haohan Wang. 2024. Robust Prompt Optimization for Defending Language Models Against Jailbreaking Attacks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=jXs6Cvpe7k

[163] Zhanpeng Zhou, Zijun Chen, Yilan Chen, Bo Zhang, and Junchi Yan. 2025. On the emergence of cross-task linearity in pretraining-finetuning paradigm. In *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*. JMLR.org, Article 2559, 31 pages.

[164] Erkang Zhu, Dong Deng, Fatemeh Nargesian, and Renée J Miller. 2019. Josie: Overlap set similarity search for finding joinable tables in data lakes. In *SIGMOD*. 847–864.

[165] Erkang Zhu, Fatemeh Nargesian, Ken Q. Pu, and Renée J. Miller. 2016. LSH Ensemble: Internet-Scale Domain Search. *Proc. VLDB Endow.* 9, 12 (2016), 1185–1196. https://doi.org/10.14778/2994509.2994534

[166] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405* (2023).

[167] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043* (2023).