Model Lakes for Cross-Model Understanding

Koyena Pal
Supervised by Professors David Bau and Renée J. Miller
Northeastern University
Boston, MA, United States
pal.k@northeastern.edu

ABSTRACT

With millions of pretrained models available, finding the right model for a task, understanding what individual models do, and characterizing how they differ from one another remain significant challenges. To address this, we present a vision for model lakes for organizing and querying large model collections. Currently, practitioners rely on manually written documentation to understand and choose models; however, not all models have complete or reliable documentation. Inspired by research on data lakes, we introduce model lakes and formalize key model lake tasks, including model attribution, versioning, search, and benchmarking, and discuss fundamental research challenges in the management of large model collections. To understand behavioral relationships across models in a lake, we present a study of cross-model chain-of-thought generalization, motivated by the search for good natural-language explanations: whether a chain-of-thought produced by one language model reliably guides other language models to the same answer, providing a measurable approximation of explanation generalizability across a model population. Given the importance of reasoning as a core model capability, we propose reasoning skill-based model characterization and recommendation as a unifying future direction for model lake research.

VLDB Workshop Reference Format:
Koyena Pal. Model Lakes for Cross-Model Understanding. VLDB 2026 Workshop: PhD Workshop.

1 INTRODUCTION

As the number of pre-trained models grows, comparing them and selecting the right one for specific tasks becomes increasingly difficult. Documentation, particularly model cards [44], aims to provide essential insights, but Liang et al. [37] 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 [39], MLCommons [43, 54], and Responsible Foundation Model Development Cheatsheet [38] 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 study and formalization of tasks for "model lakes" [50] — a system containing numerous heterogeneous pre-trained models and related data in their natural formats (one example is Hugging Face [20]). We introduce *model lakes*, as a parallel to data lakes, and discuss how important innovations in data lakes [48], including data discovery, annotation, and version management, can (and should) be applied within model lakes and studied with the same vigor.

But organization alone is insufficient: to meaningfully manage a model collection, we need to understand how models relate to each other by behavior, not just by metadata. This thesis addresses both problems: a formalization and roadmap for managing large model collections, and methods for understanding cross-model behavioral relationships. This research focuses on two threads. In the first, we formalize key model lake tasks including model attribution, versioning, search, and benchmarking, and discuss fundamental research challenges in the management of large model collections. In the second, we study cross-model chain-of-thought generalization as a concrete instance of cross-model behavioral understanding, asking whether a chain-of-thought produced by one language model reliably guides other language models to the same answer. In the following sections, Section 2 presents our model lakes framework [50]. Section 3 presents our study of cross-model CoT generalization [51]. Section 4 concludes with future work on reasoning skill-based model characterization and recommendation.

2 MODEL LAKES

In Pal et al.[50], we define a model as a tuple $\mathcal{M} = (\mathcal{D}, \mathcal{A}, f_*, \theta, p_\theta)$, where training data $\mathcal{D}$ and algorithm $\mathcal{A}$ can be traced through documentation, architecture f_* and parameters θ come from accessible model weights, and behavior p_θ from observable outputs. A model can be analyzed from three viewpoints: its **history**, **intrinsic** composition, and/or **extrinsic** behavior. The distinction is useful because in a model lake, certain aspects may be unavailable. For instance, intrinsic details of some models may be inaccessible, requiring analysis to rely solely on extrinsic observations or historical records to understand a model's behavior. We use this distinction to analyze four primary model lake tasks.

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" [8]. 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?

Proceedings of the VLDB Endowment. ISSN 2150-8097.

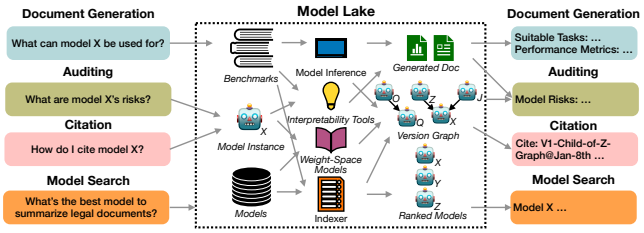

Figure 1: Model Lakes Design. A model lake stores models and processes them using techniques, like inference, interpretability, weight-space modeling and indexing to support various user interactions. It generates outputs like version graphs, model cards and ranked models, refining them into human-readable results, as shown on the figure's right side.

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 was created from other models (similar to studying how a version of a data set or table may have been derived from another [60]). One possible definition of model versioning that uses an intrinsic viewpoint [28] 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 [40]: 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.

2.1 Model Lake Research Challenges

A model lake is illustrated in Figure 1, where rather than APIs, we envision data scientists interacting with the lake using queries.

Benchmarking Benchmarking plays a crucial role in evaluating model performance and remains a well-established research area; however, benchmarks for newer topics such as model attribution and versioning are urgently needed. Broadly, there are two primary types of benchmarking: (1) evaluating how well a model approximates a ground truth distribution, and (2) assessing performance against a targeted evaluation metric. Beyond traditional performance metrics, benchmarking also considers biases related to protected attributes [56] and environmental impact [35]. Model lake benchmarks lack large-scale, publicly available datasets that mimic realistic conditions of diverse models. Preliminary efforts include Lu et al. [40], who released a benchmark for model search with 259 public image generative models, and Model Zoo [57], which is limited to vision models. Creating large-scale model lake benchmarks integrating diverse modalities remains an open challenge. An important topic is the development of lifelong benchmarks [53] that address increasingly complex scenarios as models evolve. To advance model attribution research, a benchmark dataset is needed that includes labeled model parameters, architectures, and detailed transformation records (e.g., fine-tuning, model editing), which can be extended to versioning by tracking previous and subsequent model versions.

Model Inference Deploying effective solutions for model lake tasks is challenging with respect to *usability*, *scalability*, and *computational cost*. Because model lake operations may require large-scale inference, indexing, benchmarking, and comparison across many models, they can significantly increase computational cost and energy consumption. Hence, effective solutions should also aim to minimize redundant computation and account for their energy footprint. The model inference component involves identifying appropriate benchmarks, generating relevant prompts, and selecting

suitable models to address a user query. While users can manually run prompts and select models, this is prone to errors and suboptimal outcomes, especially when users lack expertise. For example, a classifier's behavior may be misinterpreted if a user does not understand the data it was trained on. Model lake tasks can incorporate intrinsic model properties (e.g., weights) to guide users toward more accurate interpretations, and this process can be automated using an AI agent. By combining benchmarking and attribution, we can explain model behavior and output in a unified way.

Indexer A central component of a model lake is the indexer, which embeds models and provides scalable sublinear search over model embeddings using ranking functions tailored to the task. Search methods must scale to handle millions of models and adapt to newer models with advanced capabilities [64]. Indices like HNSW [42] have proven effective for nearest-neighbor search in data lakes [21], but provide no formal correctness guarantees and their use in model lakes remains under-explored. Effective model embedding is crucial for accurate comparison and ranking; Wang et al. [63] explore this but their work is not inclusive of all model types. Many model lake tasks will benefit from a hybrid approach that indexes both metadata and model embeddings — for example, model search can combine intrinsic model embeddings with search over verified model cards, while versioning can use embeddings to identify parent-child relationships and assess model distinctiveness.

Weight-Space Modeling Weight-space modeling is a hyper-representation learning approach where a neural network is trained to process the weights of other models [17], useful for distinguishing between models in complex scenarios such as model stitching. Zhou et al. [69] reveal a linear connection between fine-tuned models, a promising direction for dynamic benchmark selection by learning from previous model-dataset relationships. The primary challenge is identifying the most relevant aspects of the model for training while considering the weight-space model architecture [49] and ensuring sufficient data diversity to avoid overfitting.

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 [67], or when unlearning learned knowledge [6, 18, 22, 31, 34]. Attribution questions also apply to the source model's architecture. Approaches like circuit discovery [14, 19, 62] can help identify the computational origin of model behavior. **Model inversion** can be used to recover an input prompt given an output [29, 46, 47]. These methods are also a promising route for understanding the impact of training data on internal model states. In recent work, Sharkey et al. [58] 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.

3 CROSS-MODEL REASONING GENERALIZATION

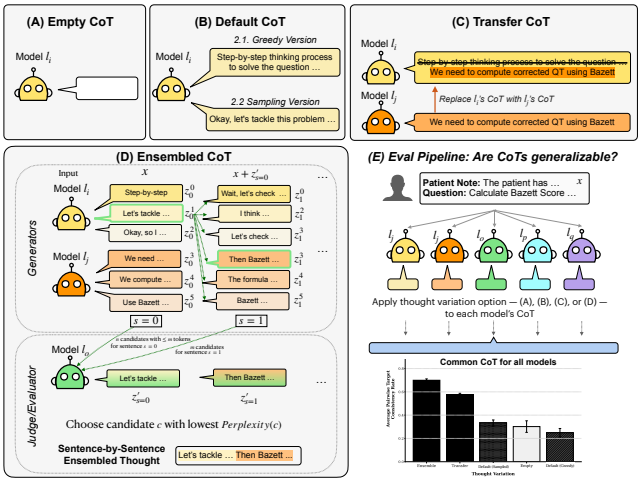

Figure 2: *Methods for eliciting chain-of-thought (CoT) explanations and evaluating them.* We evaluate explanation generation by querying LLMs for the answer to questions using CoTs in four different ways: (A) Empty CoT. No reasoning text is provided between the model's thinking tags. (B) Default CoT. The model's own reasoning is used. (2.1) uses greedy decoding, while (2.2) uses nucleus sampling. (C) Transfer CoT. Reasoning from one model is directly transferred to another, replacing its own. (D) Ensembled CoT. A generator–evaluator loop. Generator models produce $n = 3$ candidate sentences (≤ 15 tokens each), forming k candidates. These are scored by the evaluator, and the least surprising candidate (lowest perplexity) is appended to the growing ensembled thought. This updated context is fed back into the generators, and the process repeats until an end-of-thought or maximum token limit is reached. (E) After eliciting CoTs in these four settings, we evaluate their generalization to new LRMs. The transfer CoT and Ensembled CoT significantly improve the consistency of the answer produced between the model producing the original CoT and the model producing the final answer (results shown here on the MedCalc-Bench dataset).

Large reasoning models (LRMs) produce chain-of-thought (CoT) traces, which include step-by-step natural language reasoning sequences generated before producing a final answer [1, 24, 25]. These traces illustrate human-readable explanations of what the model is "thinking", but prior work has questioned the faithfulness of these explanations with respect to the model's internal decision-making process [5, 59, 66]. We examine a complementary property in our work [51]: to what extent to these CoT explanations *generalize* across models. For an explanation to be useful, it must not only be accurate but understandable, i.e., when presented to a new model, it

should lead that model to draw the same conclusions. Understanding cross-model CoT generalization is directly relevant to model lakes because if behavioral consistency across models is measurable, it provides a principled signal for characterizing and comparing models in a lake beyond metadata and benchmark scores.

Related Work A large body of work extends CoT prompting [65] by probing or refining the explanations it produces. Examples include evaluating counterfactuals introduced into the CoT [23], testing their robustness to mistakes introduced into the reasoning chain [36], or using contrastive CoT to induce reliance on the reasoning chain [13]. Prior work on consistency measures whether a model generates consistent explanations on similar examples [12, 26]. We extend this to *cross-model consistency*: whether different models converge on the same answer given the same explanation, a distinct inter-model property. Explanation evaluation frameworks [16, 30] distinguish consistency, plausibility, and faithfulness as complementary dimensions; our work introduces cross-model generalization as a fourth dimension relevant to model populations.

Approach We evaluate CoT transfer across five large reasoning models (LRMs) with distinct training pipelines on two benchmarks, namely `MedCalc-Bench` [32], a medical domain reasoning benchmark, and `Instruction Induction` [27], a general reasoning benchmark. Given a generating model l_{gen} and a problem x, we elicit a CoT explanation $z = l_{\text{gen}}(x)$ and transfer it to an evaluating model l_{eval} (see Figure 2) and then measure accuracy against ground truth and pairwise consistency across model populations. We additionally study an ensembled CoT condition where reasoning traces are constructed collaboratively across models.

Results We find that LRM explanations can generalize to other LRMs even when they induce an inaccurate answer, which suggests that CoT captures transferable reasoning structure beyond correct answers. Second, explanations that generalize across models receive higher user preference ratings, validating consistency as a proxy for explanation quality. Third, RL post-training that improves a model's performance improves its CoT consistency but not necessarily its transfer accuracy, revealing a nuanced relationship between model capability and explanation generalizability. Together, these results establish cross-model CoT consistency as a concrete, measurable behavioral signal, laying the groundwork for skill-based characterization within a model lake.

4 SKILL-BASED MODEL DISCOVERY

The notion of skill has had varied definitions, from Bloom et al. [7] defining skills as generalized modes of operation for dealing with problems, distinct from domain-specific knowledge, to Anderson and Krathwohl [3] extending this into a hierarchical taxonomy of cognitive competencies. It is further argued that skills and knowledge are intertwined, hence the term "competencies" to reflect domain-specificity [33]. These frameworks motivate capturing model capabilities as structured, hierarchical competencies instead of aggregated benchmark scores.

In the LLM literature, Arora and Goyal [4] formalize skills as a bipartite graph between latent skills and text pieces. They show that factor analytic approaches reveal that LLM performance decomposes into a small set of interpretable skill dimensions [9, 41], while skill-level analyses expose fine-grained model differences obscured by aggregate metrics [45, 61]. EvalTree [68] constructs hierarchical capability trees over model weaknesses via recursive clustering of capability annotations. Skill-based frameworks have also been applied to training data selection [11] and scaling law prediction [52].

These works characterize skills within individual models or across fixed model populations using benchmark scores. However, it remains a challenge to index, retrieve, and compare models by skill within a managed model collection. Our proposed future work looks into exploring model search where skill profiles can be queryable, storable attributes in a model lake. These would be derived from behavioral signals rather than benchmark co-performance, and organized into hierarchical skill taxonomies.

Current model retrieval research is dataset-centric where given a target dataset, users find the best performing model [10]. We propose characterizing models by *skill profiles*, showing structured representations of what a model excels at or where it could be improved. These would be derived from behavioral signals rather than benchmark scores alone.

CoT consistency explored earlier showcases one potential behavioral signal. Models that produce transferable chain-of-thought may possess a generalizable reasoning skill measurable without benchmark evaluation. Even models not built for reasoning can be prompted to produce reasoning steps, and others can be compared through their outputs on shared inputs. The open problem is extending this to a broader skill taxonomy: what behavioral signals define other skills, how do skills compose, and how stable are skill profiles across fine-tuning steps? Recent works suggest scalable ways to elicit such signals. STaD [2] scaffolds benchmark tasks to expose compositional skill gaps, and Rinberg et al. [55] show a weak model can recover much of a strong model's capability by asking it as few as ten yes/no questions about its own reasoning and revising its answer based on the responses. We can adapt these methods for the model lake by populating skill dimensions across many models. The binary-question pipeline [55] could surface which skills a model has and lacks, and scaffolded task variation [2] could then verify each candidate skill under systematic variation, which can create queryable, comparable profiles.

Given the skill profiles, we can then ask the following: given a target task described in terms of required skills, retrieve models whose skill profiles are most compatible. This can be an alternative view on looking at models where, instead of dataset performance as shown in ModelLens [10], which relies on leaderboard co-performance records and can degrade on entirely new task types with no benchmark history, we showcase by skills, drawing on structured sources like ModelTables [15] to enrich these profiles.

When a model is fine-tuned, some skills are introduced, some are evolved and some may have regressed. For instance, a model fine-tuned on medical data may gain domain knowledge while losing general reasoning skill. Hence, a skill profile for these models can help track how skills evolve across fine-tune lineages, which may enable users to ask for a base or pre-trained model responsible for the fine-tuned model's specific reasoning skill or generalizability. This connects skill-based characterization to the model lake search and attribution formalized in Section 2, and is a long-term extension of the cross-model understanding agenda.

ACKNOWLEDGMENTS

This work was supported in part by NSF award numbers IIS2107248, IIS-1956096, and IIS-2325632, by a grant from Coefficient Giving and by generous funding from the Cambridge-Boston Alignment Initiative (CBAI) Fellowship.

REFERENCES

[1] Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. gpt-oss-120b & gpt-oss-20b Model Card. *arXiv preprint arXiv:2508.10925* (2025).

[2] Sungeun An, Swanand Ravindra Kadhe, Shailja Thakur, Chad DeLuca, and Hima Patel. 2026. STaD: Scaffolded Task Design for Identifying Compositional Skill Gaps in LLMs. arXiv:2604.18177 [cs.CL] https://arxiv.org/abs/2604.18177

[3] Lorin W Anderson and David R Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition.* Addison Wesley Longman, Inc.

[4] Sanjeev Arora and Anirudh Goyal. 2023. A theory for emergence of complex skills in language models. *arXiv preprint arXiv:2307.15936* (2023).

[5] Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, et al. 2025. Chain-of-thought is not explainability. *Preprint, alphaXiv* (2025), v2.

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

[7] Benjamin S Bloom et al. 1956. Taxonomy of. *Educational Objectives* 250, 2025 (1956), 3.

[8] 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.

[9] Ryan Burnell, Han Hao, Andrew RA Conway, and Jose Hernandez Orallo. 2023. Revealing the structure of language model capabilities. *arXiv preprint arXiv:2306.10062* (2023).

[10] Rui Cai, Weijie Jacky Mo, Xiaofei Wen, Qiyao Ma, Wenhui Zhu, Xiwen Chen, Muhao Chen, and Zhe Zhao. 2026. ModelLens: Finding the Best for Your Task from Myriads of Models. arXiv:2605.07075 [cs.LG] https://arxiv.org/abs/2605.07075

[11] Mayee Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. 2023. Skill-it! a data-driven skills framework for understanding and training language models. *Advances in Neural Information Processing Systems* 36 (2023), 36000–36040.

[12] Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen McKeown. 2024. Do models explain themselves? counterfactual simulatability of natural language explanations. In *Proceedings of the 41st International Conference on Machine Learning*. 7880–7904.

[13] Yew Ken Chia, Guizhen Chen, Luu Anh Tuan, Soujanya Poria, and Lidong Bing. 2023. Contrastive Chain-of-Thought Prompting. arXiv:2311.09277 [cs.CL]

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

[15] Zhengyuan Dong, Victor Zhong, and Renée J Miller. 2025. ModelTables: A Corpus of Tables about Models. *arXiv preprint arXiv:2512.16106* (2025).

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

[17] 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.

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

[19] 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.

[20] Hugging Face. 2026. https://huggingface.co

[21] 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.

[22] 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*.

[23] Yair Gat, Nitay Calderon, Amir Feder, Alexander Chapanin, Amit Sharma, and Roi Reichart. 2023. Faithful Explanations of Black-box NLP Models Using LLM-generated Counterfactuals. *arXiv preprint arXiv:2310.00603* (2023).

[24] Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. 2025. OpenThoughts: Data Recipes for Reasoning Models. *arXiv preprint arXiv:2506.04178* (2025).

[25] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).

[26] Peter Hase and Mohit Bansal. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?. In *Proceedings of the Association for Computational Linguistics*. https://aclanthology.org/2020.acl-main.491

[27] Or Honovich, Uri Shaham, Samuel R Bowman, and Omer Levy. 2022. Instruction Induction: From Few Examples to Natural Language Task Descriptions. *arXiv preprint arXiv:2205.10782* (2022).

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

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

[30] Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?. In *Proceedings of the Association for Computational Linguistics*. https://aclanthology.org/2020.acl-main.386

[31] 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

[32] Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, Soren Dunn, Serina Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad Safranek, Abid Anwar, Andrew Zhang, et al. 2024. Medcalc-bench: Evaluating large language models for medical calculations. *Advances in Neural Information Processing Systems* 37 (2024), 84730–84745.

[33] Developing Transferable Knowledge. 2012. Education for life and work. (2012).

[34] 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)*.

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

[36] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702* (2023).

[37] 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:2402.05160 [cs.SE]

[38] 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

[39] 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

[40] 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

[41] Aviya Maimon, Amir DN Cohen, Gal Vishne, Shauli Ravfogel, and Reut Tsarfaty. 2025. IQ Test for LLMs: An Evaluation Framework for Uncovering Core Skills in LLMs. *arXiv preprint arXiv:2507.20208* (2025).

[42] 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.

[43] 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

[44] 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* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

[45] Mazda Moayeri, Vidhisha Balachandran, Varun Chandrasekaran, Safoora Yousefi, Thomas Fel, Soheil Feizi, Besmira Nushi, Neel Joshi, and Vibhav Vineet. 2024. Unearthing skill-level insights for understanding trade-offs of foundation models. *arXiv preprint arXiv:2410.13826* (2024).

[46] 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

[47] 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

[48] 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.

[49] 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

[50] Koyena Pal, David Bau, and Renée J. Miller. 2025. Model Lakes. In *Proceedings 28th International Conference on Extending Database Technology, EDBT 2025, Barcelona, Spain, March 25-28, 2025*. OpenProceedings.org, 985–995. https://doi.org/10.48786/EDBT.2025.81

[51] Koyena Pal, David Bau, and Chandan Singh. 2026. Do explanations generalize across large reasoning models? arXiv:2601.11517 [cs.CL] https://arxiv.org/abs/2601.11517

[52] Felipe Maia Polo, Seamus Somerstep, Leshem Choshen, Yuekai Sun, and Mikhail Yurochkin. 2024. Sloth: scaling laws for LLM skills to predict multi-benchmark performance across families. *arXiv preprint arXiv:2412.06540* (2024).

[53] 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:2402.19472 [cs.LG] https://arxiv.org/abs/2402.19472

[54] 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

[55] Roy Rinberg, Annabelle Michael Carrell, Simon Henniger, Nicholas Carlini, and Keri Warr. 2026. Haiku to Opus in Just 10 bits: LLMs Unlock Large Compression Gains. arXiv:2604.02343 [cs.LG] https://arxiv.org/abs/2604.02343

[56] 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.

[57] 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.

[58] 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:2501.16496 [cs.LG] https://arxiv.org/abs/2501.16496

[59] Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941* (2025).

[60] 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

[61] Lindia Tjuatja and Graham Neubig. 2025. BehaviorBox: Automated discovery of fine-grained performance differences between language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 18851–18873.

[62] 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*.

[63] 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

[64] 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.

[65] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[66] Zidi Xiong, Shan Chen, Zhenting Qi, and Himabindu Lakkaraju. 2025. Measuring the Faithfulness of Thinking Drafts in Large Reasoning Models. *arXiv preprint arXiv:2505.13774* (2025).

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

[68] Zhiyuan Zeng, Yizhong Wang, Hannaneh Hajishirzi, and Pang Wei Koh. 2025. Evaltree: Profiling language model weaknesses via hierarchical capability trees. *arXiv preprint arXiv:2503.08893* (2025).

[69] 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* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 2559, 31 pages.