Open Technical Problems in Open-Weight AI Model Risk Management

Stephen Casper *MIT CSAIL* *scasper@mit.edu*
Kyle O'Brien *ERA Fellowship*
Shayne Longpre *MIT*
Elizabeth Seger *Demos*
Kevin Klyman *Stanford University*
Rishi Bommasani *Stanford University*
Aniruddha Nrusimha *MIT CSAIL*
Ilia Shumailov *AI Sequrity Company*
Sören Mindermann *Mila, Université de Montréal LawZero*
Steven Basart *Center for AI Safety*
Frank Rudzicz *Dalhousie University Vector Institute*
Kellin Pelrine *FAR.AI*
Avijit Ghosh *Hugging Face*
Andrew Strait *UK AI Security Institute*
Robert Kirk *UK AI Security Institute*
Dan Hendrycks *Center for AI Safety*
Peter Henderson *Princeton University*
Zico Kolter *Carnegie Mellon University*
Geoffrey Irving *UK AI Security Institute*
Yarin Gal *UK AI Security Institute OATML, University of Oxford*
Yoshua Bengio *Mila, Université de Montréal*
Dylan Hadfield-Menell *MIT CSAIL*
Reviewed on OpenReview: *https://openreview.net/forum?id=8QyGLnFkzc*

Abstract

Frontier AI models with openly available weights are steadily becoming more powerful and widely adopted. However, compared to proprietary models, open-weight models pose different opportunities and challenges for effective risk management. For example, they allow for more open research and testing. However, managing their risks is also challenging because they can be modified arbitrarily, used without oversight, and spread irreversibly. Currently, there is limited research on safety tooling specific to open-weight models. Addressing these gaps will be key to both realizing their benefits and mitigating their harms. In this paper, we present 16 open technical challenges for open-weight model safety involving training data, training algorithms, evaluations, deployment, and ecosystem monitoring. We conclude by discussing the nascent state of the field, emphasizing that openness about research, methods, and evaluations – not just weights – will be key to building a rigorous science of open-weight model risk management.

1 Introduction

Open-weight AI models – models whose weights are publicly available to download – have quickly grown in their capabilities and prominence (Cottier, 2024; Bhandari et al., 2025). 2025 has been a major year for advanced open language, image, and video models (see Table 2 for examples). Simultaneously, proprietary model developers have reported that their models are approaching critical risk thresholds (Google AI, 2025; Anthropic, 2025; OpenAI, 2025). Meanwhile current work estimates that the capabilities of frontier open-weight models only lag 6 to 12 months behind (Cottier, 2024; Maslej et al., 2025). This suggests that open-weight models could soon cross critical capability thresholds.

Open- versus closed-weight deployments come with different safety tradeoffs. While open-weight models allow for more open research and testing, they also come with a greater potential for misuse. Frontier closed weight model developers often rely on a complex combination of training interventions, content classifiers, and enforcement of acceptable use policies to reduce risks (e.g., Google AI, 2025; Anthropic, 2025; OpenAI, 2025). However, none of these methods provide reliable assurances for open-weight models, which can be used, tampered with, and distributed without limitations. Compared to closed-weight models, the attack surface for open-weight models is larger, and the toolkit of reliable techniques for defending them is less well studied. Furthermore, as increasing numbers of open models are released and shared (Bhandari et al., 2025), it is difficult to understand the extent of their spread, usage, and impacts.

Building the field of technical safeguards for open models will be key to capturing their benefits and minimizing their risks. In Section 2, we expand on difficulties, highlighting how *tampering threats* and the *diffuse spread of open models* are defining challenges for managing their risks. Section 3 then outlines technical objectives targeting these gaps. We organize them into five categories spanning the model lifecycle. Objectives 1-3 address threats to open-weight models from harmful tampering, while 3-5 focus on improving access to actionable information about the real-world uses and risks of open models:

1. *Training data curation* methods for preventing models from learning harmful capabilities (Section 4.1). Recent research has shown that these methods are effective at making open-weight models resist few-shot learning harmful behaviors.
2. *Tamper-resistant training and 'unlearning' algorithms* for building additional defenses against malicious fine-tuning and other forms of tampering (Section 4.2).
3. *Model tampering evaluations* for testing model risks under misuse threats from harmful tampering (Section 4.3). These methods are necessary to evaluate real-world risks from downstream model modifications to open-weight models.
4. *Staged deployment strategies* which allow developers to experiment with partial access before a full open release (Section 4.4). These techniques allow developers to monitor for unexpected uses and modify their plans for safeguards and release before a model is made openly available.
5. *Model provenance and forensics* strategies for monitoring real-world uses and impacts (Section 4.5). These strategies offer tools for developers, academics, and other stakeholders to study the diffuse open-weight model ecosystem.

2 Why is open-weight model risk management challenging?

Open-weight models can be used and adapted widely without centralized control. This is key to their benefits, enabling more widespread research and decentralization of power (Bommasani et al., 2024; Telecommunications & Administration, 2024; Seger, 2024; Eiras et al., 2024b; Kapoor et al., 2024; François et al., 2025; Longpre et al., 2025b; Miller et al., 2025). For example, the release of DeepSeek R1 has enabled independent safety research of near-frontier models (Goodfire, 2025; Zhou et al., 2025a). However, these same characteristics also make open-weight model risk management distinctly challenging.

Users can disable safety tools that are external to the model. A key strategy for managing frontier AI risks is to augment models with external safeguards to monitor for signs of risk and intervene to prevent

harm (Bengio et al., 2025b;a; Sharma et al., 2025; Korbak et al., 2025). For example, it is common for AI models to be deployed with input and/or output filters designed to detect and block harmful uses. These kinds of tools can be very valuable to release alongside open-weight models, but they are also trivial for users with access to the model to disable.

Downstream users can *tamper* with open models via fine-tuning or other modifications to remove safeguards or add harmful capabilities. Open and closed-weight AI models alike can both be vulnerable to jailbreaks or other adversarial prompts that elicit harmful behavior. However, open-weight models are additionally prone to powerful tampering threats. Benign fine-tuning, adversarial fine-tuning, and other modifications to a model have been shown to effectively elicit harmful behaviors and capabilities – even from models that are relatively safe off the shelf (e.g., Qi et al., 2023; Hu et al., 2025b; Greenblatt et al., 2024; Wei et al., 2024; Hofstätter et al., 2025; Che et al., 2025).[1] There is growing precedent for open-weight models being fine-tuned and shared specifically for harmful uses. For example, modified open-weight diffusion models have become the most common tools used for creating synthetic child sexual abuse material (IWF, 2024; Hawkins et al., 2025; Vaughan). Recent research has also identified open-weight model variants that have been fine-tuned specifically to perform malicious tasks (Simonovich, 2025). Meanwhile, thousands of open-weight text models have been specifically fine-tuned to disable safeguards. With enough fine-tuning on enough data, safeguards for any model can be undone, meaning that practical anti-tampering techniques can only hope to make harmful forms of fine-tuning sufficiently onerous.

Open-weight models can be spread quickly and irreversibly. If a closed-weight model is found to pose hazards, risk-conscious developers can add patches or pull the model from distribution. Consider, for example, OpenAI's April 2025 update of GPT-4o. After release, external evaluation identified excessive sycophancy and encouragement of self-harm. In response, OpenAI reverted to a previous version of the model (OpenAI, 2025b). In contrast, OpenAI's open-weight release of gpt-oss-120b, which currently has over 3 million monthly downloads from HuggingFace, was not reversible. While ceasing service to a model can make it much less accessible (e.g., Solaiman et al., 2025; Seger, 2024), there is no reliable way to prevent existing copies of the model from being used and shared.

Open-weight models cannot be centrally monitored or moderated. When closed-weight models are released, they are generally made available through an API controlled by the model deployer. This allows for the developer to use 'Know Your Customer' strategies (Jami Pour et al., 2024), monitor for misuse (OpenAI, 2025a; Yueh-Han et al., 2025; Brown et al., 2025), and enforce acceptable usage policies. In contrast, open-weight models generally cannot be centrally monitored.

Open-weight models have more complex supply chains than closed models. Model supply chains involve many resources, including talent, data, compute, and infrastructure (Cen et al., 2023; Longpre et al., 2023a). Closed-weight models can be developed by multiple actors. However, they tend to be developed in a relatively centralized and coordinated way. In contrast, open-weight models more often result from many stages of modification and redistribution, often across jurisdictions. Models with complex supply chains are more prone to single actors introducing harmful behaviors (including backdoors, Hanif et al., 2025), and they make it more difficult to determine accountability for harms (Nissenbaum, 1996; Cooper et al., 2022).

It can be difficult to track the real-world spread, usage, and impacts of open-weight models. Because their usage is typically much less centralized, it is often hard for researchers to thoroughly understand the spread, uses, and impact of open-weight models. This makes it more difficult to study risks, perform cost-benefit analysis, and identify effective points of intervention in the open-weight model ecosystem.

3 The Toolkit

3.1 Technical Safeguards for Open-Weight models

Scope and relation to prior work: This paper builds on prior research on open-weight models and their implications for risk management. This includes prior work on outlining the risks and benefits of highly capable open-weight models (Chan et al., 2023; Seger et al., 2023; Eiras et al., 2024b; Kapoor et al., 2024;

[1]Notably, several of these papers cited here demonstrated model vulnerabilities to tampering threats via fine-tuning APIs.

Bengio et al., 2025b;b), proposals for risk management frameworks (Liang et al., 2022; Longpre et al., 2025a; Gal & Casper, 2025), and considerations for governance (Bommasani et al., 2024; Telecommunications & Administration, 2024). In particular, our work is closely related to Seger (2024), François et al. (2025), Srikumar et al. (2024), and UK AISI (2025), which each taxonomize and overview approaches for managing risks from open-weight AI models. To complement these past works, **we focus only on open problems for technical safeguards[2] that have distinct implications[3] for open-weight models.**

Taxonomizing technical safeguards with distinct applicability to open-weight model risk management: Based in part on taxonomies provided in Seger (2024), François et al. (2025), Srikumar et al. (2024), and UK AISI (2025), we organize technical safeguards with distinct implications for open-weight model risk management in Table 1. Our taxonomy includes five categories corresponding to different stages of a model's lifecycle: training data curation, training, evaluation, deployment, and post-deployment monitoring.

Approach	Seger (2024)	François et al. (2025)	Srikumar et al. (2024)	UK AISI (2025)
Training Data Curation (Section 4.1)	✓	✓	✓	✓
Tamper-Resistant Training & **Unlearning Algs.** (Section 4.2)	✓	✓	✓	✓
Model Tampering Evaluations (Section 4.3)	✓	✗	✓	✓
Staged Deployment Strategies (Section 4.4)	✓	✗	✓	✓
Model Provenance & **Forensics** (Section 4.5)	✓	✗	✓	✓

Table 1: Our taxonomy of technical safeguards with distinct implications for open-weight models sorts methods into five categories corresponding to the different stages of a model's lifecycle: training data curation, training, evaluation, deployment, and post-deployment monitoring. In this table, we overview these methods' coverage in prior public work.

3.2 What this Paper Does Not Cover

The primary goal of this paper is to help build the technical science of open-weight model risk management. As a result, we only focus on open problems for technical safeguards with distinct implications for open-weight models. However, this is not to say that other risk-management strategies are not crucial for managing risks from open-weight models. For the purposes of this paper, the following strategies discussed by Bommasani et al. (2024), Telecommunications & Administration (2024), Seger (2024), François et al. (2025), Srikumar et al. (2024), and UK AISI (2025) are out of scope:

- Technical tools for AI risk management that do not have distinct implications for open-weight models:
  - General (non-tamper-resistant) safety fine-tuning techniques (Seger, 2024; François et al., 2025; Srikumar et al., 2024; UK AISI, 2025)
  - Safety scaffolding and content moderation tools (Seger, 2024; François et al., 2025; Srikumar et al., 2024; UK AISI, 2025)
  - Rigorous black-box evaluations (Seger, 2024; François et al., 2025; Srikumar et al., 2024; Zhu et al., 2025c)

[2]We define a "technical safeguard" as a technique which uses methods from machine learning research.

[3]We consider a safeguard to have "distinct implications" for open-weight models if it is applicable for open-weight releases and cannot be trivially disabled. For example, we do not consider filters for harmful outputs to have disinct implications for open-weight model risk management as they are also used for closed-weight models and can be trivially disabled in open-weight models.

- Data provenance/forensics (Bommasani et al., 2024; Seger, 2024; François et al., 2025; Srikumar et al., 2024; UK AISI, 2025; Longpre et al., 2024b)
- Monitoring fine-tuning APIs (see also Halawi et al., 2024; Davies et al., 2025)

- Nontechnical approaches for open-weight model risk management:
  - Having an acceptable use policy and proactively enforcing violations (Seger, 2024; Srikumar et al., 2024)
  - Transparency and documentation of code, methods, and evaluation results (Seger, 2024; Srikumar et al., 2024; UK AISI, 2025)
  - Know-your-customer methods (Seger, 2024; UK AISI, 2025) (see also Jami Pour et al., 2024)
  - Implementing incident reporting policies and infrastructure (Seger, 2024; Srikumar et al., 2024) (see also Cattell et al., 2024; McGregor et al., 2025; Longpre et al., 2025b)
  - Monitoring misuse, unintended uses, and user feedback (Telecommunications & Administration, 2024; Seger, 2024; Srikumar et al., 2024)
  - Pulling and replacing models found to pose hazards (Seger, 2024; UK AISI, 2025)

- Accelerating progress in the development of open-weight models with beneficial societal impacts (Bommasani et al., 2024; Seger, 2024; Telecommunications & Administration, 2024; François et al., 2025)

- Governance strategies that carry distinct implications for the open-weight model ecosystem (Bommasani et al., 2024; Seger, 2024; Telecommunications & Administration, 2024; François et al., 2025)

4 Open Technical Problems

Here, following the taxonomy in Table 1 above, we discuss open technical problems in open-weight model risk management spanning the model lifecycle from pretraining to post-deployment ecosystem monitoring. The first three challenges we discuss pertain to open-weight model resistance to harmful fine-tuning and other forms of tampering. Section 4.1 focuses on *training data* curation and its implications for how well a final model resists harmful tampering. Section 4.2 additionally covers *training* algorithms to make models resistant to tampering. Section 4.3 focuses on *evaluating* open-weight models under realistic tampering threats. The final two challenges pertain to how models are released and monitored after they are developed. Section 4.4 discusses strategies for *deploying* models in stages. Finally, Section 4.5 covers tools to facilitate the *post-deployment monitoring* of model uses and impacts in the open-weight ecosystem

4.1 Training Data Curation

Training data curation is increasingly understood as a key intervention for improving model safety both off-the-shelf and under tampering. Frontier AI models are prone to learning harmful information during training (e.g., Phuong et al., 2024; Anthropic, 2025). An intuitive countermeasure is to prevent models from learning harmful capabilities by minimizing exposure to unsafe training data. This aligns with evidence that models acquire most of their core knowledge during pretraining (Zhou et al., 2023; Raghavendra et al., 2024; Chang et al., 2024), where they are exposed to a diverse corpus of data spanning trillions of tokens (e.g. Yang et al., 2025; Meta AI, 2025; Agarwal et al., 2025). However, once knowledge has been internalized by a model, it is empirically difficult to remove or substantially modify (Jia et al., 2021; Anwar et al., 2024). This suggests that training data interventions, particularly during pretraining, could have the potential to shape the core knowledge, concepts, and propensities of the model towards safer outcomes.

Existing work: Curating training data at scale and filtering harmful content (such as text with instructions for performing illegal actions or images/videos depicting child sexual abuse) are widely understood as a key means of training safer AI models (e.g., Nichol et al., 2022; Longpre et al., 2023b; Thorn & All Tech Is Human, 2024; Seger, 2024; Longpre et al., 2025a; Liu et al., 2024c; François et al., 2025; Srikumar et al., 2024). Blocking data from known risky sources, such as websites with a high prevalence of adult and toxic

content, is a common but not standardized practice in dataset design (Soldaini et al., 2024; Penedo et al., 2024; Li et al., 2024a). Several works on open datasets have made significant contributions by releasing their data curation pipelines (Gao et al., 2020; Laurençon et al., 2023; Raffel et al., 2019; Soldaini et al., 2024; Kandpal et al., 2025). Other recent works have begun to study how pretraining data curation can be used to prevent unsafe capabilities (Korbak et al., 2023; Chen et al., 2025c; UK AISI, 2025; O'Brien et al., 2025; Wallace et al., 2025; Lee et al., 2025; Liu et al., 2025b). These works highlight notable successes, limitations, and open questions.

4.1.1. How does data curation's effectiveness differ across harm categories? Recent work has shown both successes (Maini et al., 2025; Lee et al., 2025; Chen et al., 2025c; O'Brien et al., 2025; Albalak et al.; Liu et al., 2025b) and limitations (Li et al., 2025b; Wallace et al., 2025; Wei et al., 2025) of data curation in scoping model capabilities. For instance, recent works have demonstrated pretraining filtering's ability to significantly reduce models' knowledge of biorisk-related topics (Chen et al., 2025c; O'Brien et al., 2025). However, Wallace et al. (2025) found an implementation of filtering to be ineffective when applied to gpt-oss. These findings suggest that filtering has the potential to be effective in preventing technical capability, but that it may be sensitive to implementational details. However, precise comparisons are challenging due to model cards often lacking information, such as the amount of data filtered or the amount of compute spent on filtering. Taken together, current works suggest that filtering data related to entire science or engineering domains can build more durable safeguards into models (Lee et al., 2025; O'Brien et al., 2025), while filtering data related to simpler propensities (such as toxicity or refusal of harmful requests) (Maini et al., 2025; Li et al., 2025b) or more niche science topics (Wei et al., 2025) does not. Open questions remain regarding data filtering's ability to limit potentially unsafe capabilities that are closely intertwined with beneficial capabilities, such as offensive hacking and defensive cybersecurity (Barez et al., 2025).

4.1.2. How can scaling data curation be scaled across languages, modalities, and data/model sizes? Training data curation at scale is deceptively difficult (Paullada et al., 2021) due to costs (Ngo et al., 2021), filtering errors (Ziegler et al., 2022), degradation of dataset quality (Welbl et al., 2021), the massively multilingual nature of internet text (Kreutzer et al., 2022), biases in content moderation (Welbl et al., 2021; Dodge et al., 2021; Xu et al., 2021; Stranisci & Hardmeier, 2025), and the inherently contextual nature of harmfulness (Lindner & El-Assady, 2022). When curating internet-scale datasets, efficiency, precision, and recall are crucial. Slow feedback loops exacerbate these challenges; ineffective data curation may only become apparent at the end of long training runs, potentially requiring retraining the model from scratch. Regarding efficiency and precision, O'Brien et al. (2025) recently introduced a multi-stage approach to filtering that required less than 1% of the subsequent model's training compute. However, their approach sacrificed efficiency for precision, resulting in many benign documents being filtered. Regarding recall, it is unclear whether larger and more sample-efficient models require increasingly extensive filtering. Developing frameworks that expand the pareto frontier of efficiency, precision, and recall will be key to making training data filtering more competitive (e.g., Chen et al., 2025c).

4.1.3. What is the relationship between training data contents and emergent model capabilities? More broadly, the general relationship between model architecture, the content of training data, and emergent capabilities is unclear. Recent work on influence functions (e.g., Grosse et al., 2023), out-of-context reasoning (e.g., Berglund et al., 2023; Treutlein et al., 2024; Hu et al., 2025a), coreset analysis (Pal et al., 2025), and domain-aware scaling (Hamidieh et al., 2025) has suggested a surprising ability of language models to infer generalizable knowledge from constitutive information in training data. This presents important questions. Can unsafe capabilities emerge from benign data approved by data filtering pipelines in practical settings? Relatedly, can filtering data from one domain have substantial unintended effects on model capabilities in another domain? Ultimately, a predictive and practically applicable theory of emergent capabilities in state-of-the-art models remains elusive (Wei et al., 2022; Schaeffer et al., 2025). One approach for continued work can be to study how specific behaviors emerge in simple settings. However, the most directly risk-relevant research will be empirical work to examine how realistic interventions on training data affect frontier models. In particular, understanding if and when models can learn dangerous capabilities from innocuous training data, how much data is required, what quality of data is needed, and how dynamics change with scale will be relevant and actionable for managing risks.

4.2 Tamper-Resistant Training and Unlearning Algorithms

Training algorithms designed to make models resist tampering can further improve safety under harmful modifications. Aside from training data interventions, there is also a growing body of research focused on post-training defenses for open-weight models. In particular, post-training safeguards designed to resist downstream 'tampering' modifications are a core strategy for mitigating the risks from malicious or negligent downstream use of open-weight models.

Existing work: Currently, researchers study safety fine-tuning (e.g., Bai et al., 2022) and "machine unlearning" methods (Gao et al., 2024; Liu et al., 2024b; Barez et al., 2025) as strategies for making models more strongly resist harmful behaviors uses such as assisting a user in illegal activity. However, state-of-the-art fine-tuning and unlearning algorithms have consistently been vulnerable to being undone within dozens of steps of adversarial fine-tuning. While research on defenses often reports model resistance to thousands or tens of thousands of examples of adversarial fine-tuning, to the best of our knowledge, the state of the art for tamper-resistance, as assessed by second-party red-teaming efforts, is only around several hundred steps of adversarial fine-tuning (Qi et al., 2023; Yang et al., 2023; Bhardwaj & Poria, 2023; Li et al., 2024b; Lynch et al., 2024; Huang et al., 2024b; Hu et al., 2025b; Łucki et al., 2025; Peng et al., 2024; Deeb & Roger, 2025; Qi et al., 2024; Che et al., 2025; Dorna et al., 2025). This applies even to methods that have been designed to confer tamper resistance (Łucki et al., 2025; Qi et al., 2024; Che et al., 2025). When techniques do withstand multiple rounds of supervised fine-tuning, it tends to come with major tradeoffs to a model's general knowledge and fluency (Qi et al., 2024; Zhou et al., 2024).

4.2.1. How do we develop more tamper-resistant unlearning algorithms? The persistent struggles of tamper-resistant unlearning methods prompt a reassessment of current approaches. Prior methods have involved pruning (Lo et al., 2024; Chapagain et al., 2025), meta-learning (Abdalla et al., 2025; Anonymous, 2025a; Rosati et al., 2025; Perin et al., 2025; Li et al., 2025a; Wang et al., 2025b; Yi et al., 2025), training with specialized objectives (Cao, 2025; Feng et al., 2025), training under tampering (Henderson et al., 2023; Huang et al., 2024c;a; Zheng & Yeh, 2024; Fan et al., 2025a; Cheng et al., 2025a; Zheng et al., 2025; Liu et al., 2025a; Sheshadri et al., 2025; Tamirisa et al., 2025; Sanyal et al., 2025), specially-parameterized updates (Sondej & Yang, 2025), and activation noising (Rosati et al., 2024; Pan et al., 2024; Zou et al., 2024; Tamirisa et al., 2025; Abdalla et al., 2025). Benchmarking work has yet to thoroughly compare all of these types of methods. There may be several opportunities for algorithmic innovation. Some inspiration can be taken from some successes of pretraining data filtering (O'Brien et al., 2025; Liu et al., 2025b). It is possible that running tamper-resistance algorithms for a long time and/or during pretraining could confer stronger tamper resistance. Alternative approaches could attempt to leverage a mechanistic understanding of models that are entirely ignorant about topics (e.g., O'Brien et al., 2025; Liu et al., 2025b) to design more principled training objectives compared to existing ones (e.g. Rosati et al., 2024; Zou et al., 2024; Tamirisa et al., 2025; Abdalla et al., 2025). Finally, while Huang et al. (2024c), Huang et al. (2024a), Sheshadri et al. (2025), Tamirisa et al. (2025), and others have used adversarial methods for tamper resistance, they each only train against a narrow class of tampering attacks. Training models against a more diverse assortment of tampering threats might be able to confer more generalizable tamper resistance. On the other hand, there also might be fundamental limitations for post-training methods' abilities to deeply remove or make inaccessible some types of unwanted knowledge from models. Finally, to ensure the competitiveness of safer models, future research on tamper-resistant fine-tuning will need to prioritize striking a precise balance between the removal of harmful capabilities and degrading benign ones.

4.2.2. How can we robustly edit model beliefs with minimal side effects? In contrast to making models ignorant about potentially harmful topics, some researchers have proposed introducing specific incorrect beliefs into language models about hazardous procedures (e.g., Wang et al., 2025a) to prevent the model from generating harmful outputs. For example, teaching a model incorrect information about how to acquire child sexual abuse material could be a complementary approach to both refusal training and unlearning. Belief revision can occur via specific edits to their parameters (e.g., Meng et al., 2023; Geva et al., 2023; Zhang et al., 2024) or fine-tuning (e.g., Wang et al., 2025a; Slocum et al., 2025). However, both approaches currently suffer from challenges of robust generalization (e.g., Wu et al., 2025; Zhong et al., 2024; Slocum et al., 2025), scalability (O'Brien et al., 2025), interference with other interventions (Kolbeinsson et al., 2024), and the ripple effects of belief modification (Cohen et al., 2024; Hase et al., 2024). This sug-

gests useful opportunities to develop benchmarks for surgical knowledge revision, improve scalability, limit side-effects, and demonstrate realistic use cases for mitigating specific risks with knowledge editing.

4.2.3 Can we develop models that effectively resist retrieving harmful information? LLMs do not necessarily need to know harmful information to provide it to a user. Models are increasingly being augmented with tools to search, retrieve, and synthesize information from the web (OpenAI, 2025; He et al., 2025). For example, O'Brien et al. (2025) showed that a biothreat-ignorant LLM could still effectively answer biothreat-related questions when given information with the answer in context, such as a textbook or scientific paper. This poses a unique challenge for open-weight model risk management because the standard defenses of refusal, API monitoring, and intervention can be disabled for open-weight models. One open challenge is to practically study the differences between the capabilities of domain-ignorant and domain-competent retrieval-augmented models on complex real-world tasks. Human domain experts are more effective than nonexperts at searching for answers to domain questions on the web, so it is intuitive that the same may apply for language models. Nonetheless, to our knowledge, this has not been directly tested in large models. A second challenge is to develop tamper-resistant safeguards that can defend against tool-augmentation attacks. For example, O'Brien et al. (2025) found that some machine unlearning techniques (Zou et al., 2024) were an effective off-the-shelf defense but were not tamper-resistant. Currently, tamper-resistant safeguards against these attacks remain unaddressed.

4.3 Model Tampering Evaluations

Evaluating models under tampering threats is necessary to assess real-world risks from open models. Internal and external evaluations of frontier AI models are central to emerging AI governance and risk management frameworks. Because many actors could fine-tune open-weight models with unsafe data (Huang et al., 2024b) or insert backdoors after initial pretraining (Bai et al., 2024; Zhou et al., 2025b), evaluations under these types of threats are necessary to fully assess practical risks.

Existing work: Fully assessing the risks of open-weight models requires evaluating them under "tampering" (Gal, 2024; Casper et al., 2024; Che et al., 2025; Wallace et al., 2025; O'Brien et al., 2025) threats from fine-tuning, steering, model editing, pruning, or other interventions. However, many current assessments neglect this possibility. For example, tampering evaluations are not reported on in the technical reports for most frontier open-weight models (see Table 2). Standard procedures to assess these gaps have not yet been established. For example, to our knowledge, gpt-oss has been the only frontier open-weight model where pre-release adversarial fine-tuning evaluations have been reported (Wallace et al., 2025; Agarwal et al., 2025). The current lack of common tampering evaluations creates a risk of both missing harmful uplift potential and incentivizing developers to game evaluations with superficial safeguards.

4.3.1. How can we develop rigorous benchmarking and evaluation frameworks? While it is widely understood that the potential risks from open-weight models depend greatly on how easily they can be harmfully tampered with, little tampering evaluation infrastructure exists. Notably, two recent toolkits (Hossain et al., 2025; Dombrowski et al., 2025) have introduced frameworks to evaluate the capabilities of models under a suite of tampering threats. However, they do not fully address challenges with these evaluations, such as sensitivity to different forms of elicitation (Fan et al., 2025b), hyperparameter sensitivity (Qi et al., 2024), the diversity of adversarial attacks (Łucki et al., 2025), or the existence of multiple metrics for measuring tampering attacks (tokens, steps, compute, effort, etc). No framework has yet achieved a degree of threat coverage comparable to the full model tampering toolkit, leading to patchwork evaluations in the field that can be difficult to compare and trust (Huang et al., 2024b; Hossain et al., 2025; Qi et al., 2024; Łucki et al., 2025).

4.3.2. What modifications should be used for worst-case risk estimation under model tampering? Current research on evaluating tamper resistance has principally focused on fine-tuning threats (Qi et al., 2023; Yang et al., 2023; Bhardwaj & Poria, 2023; Li et al., 2024b; Lynch et al., 2024; Huang et al., 2024b; Hu et al., 2025b; Łucki et al., 2025; Peng et al., 2024; Deeb & Roger, 2025; Qi et al., 2024; Che et al., 2025). However, other types of interventions have been known to impair the safety of models, including pruning (Wei et al., 2024), low-rank modifications (Wei et al., 2024), latent-space attacks (Bailey et al., 2025), model merging (Hammoud et al., 2024), quantization (Egashira et al., 2024; Chen et al., 2025a), distillation

(Yang et al., 2024; Angell et al., 2025), and backdoor insertion algorithms (Bai et al., 2024; Zhou et al., 2025b). These threats have not yet been studied adversarially alongside tamper-resistant algorithms. More rigorous evaluations of open-weight model risks will require considering the full tampering toolkit for model tampering. In particular, it will be important to study the extent to which simple modifications to models might be able to greatly alter their capabilities and associated risks. For example, few-shot fine-tuning and iterative reasoning have been shown to significantly improve over a model's advertised capabilities under default evaluation conditions (Muennighoff et al., 2025). Despite posing significant risk and uncertainties, these types of "capability overhangs" are not well understood for open-weight models. Finally, it is not even currently clear whether it is possible to build enough tampering resistance into models to impose meaningful barriers to misuse. Even if models resist thousands of steps of fine-tuning (O'Brien et al., 2025; Liu et al., 2025b), performing these tampering attacks may only take minutes and cost tens of dollars. It is unclear the extent to which obtaining training data can serve as a meaningful bottleneck.

4.3.3. How can we systematically identify effective attacks and defenses? Exhaustively testing tampering attacks is computationally prohibitive. For instance, even a standard fine-tuning attack can vary in multiple aspects: learning rate, training steps, training algorithm, etc. In particular, a number of recent works have shown how fine-tuning dataset contents has a large impact on the effectiveness of adversarial fine-tuning (Shen et al., 2024; Hsiung et al., 2025; Eiras et al., 2024a; PROJECTION, 2025; Xiao et al., 2025; Hu et al., 2025c; Anonymous, 2025b; Ham et al., 2025; Chen et al., 2025b; He et al., 2024). Safety measures that appear robust to some tampering threats can fail against others (Qi et al., 2024; Łucki et al., 2025; Che et al., 2025). Currently, there is not yet a general understanding of which attack configurations are necessary to stress-test safety and which are redundant. Answering this would enable more rigorous evaluation with limited computational resources. Future work on the standardized assessment of a large number of attack and defense configurations could reveal crucial patterns to guide the development of future safety approaches.

4.3.4. How can we scalably evaluate thousands of models? A major challenge to better understanding the open-weight ecosystem stems from the sheer number of existing models. Coordinated efforts to evaluate their safety properties at scale could improve practical risk management and future risk modeling. For example, platforms like Hugging Face which host and distribute large numbers of AI models can struggle to reliably identify and remove ones that violate their content policies (e.g., Maiberg, 2025). However, ecosystem-level evaluation is complicated by scale, architectural diversity, and the continuous introduction of new models. Evaluations involving tampering attacks can be particularly challenging due to the computational costs of fine-tuning and other tampering algorithms. There is a need for infrastructure for evaluating models at scale that balances efficiency with thoroughness. These approaches might also integrate new technical resources like model provenance techniques (see Section 4.5).

4.4 Staged Deployment Strategies

Gradually increasing access to a model before a full open release helps developers monitor for risks and adjust their safeguards and deployment strategies. Release strategies for AI systems do not fall into a binary between fully closed and fully open. Different strategies can strike different tradeoffs between open access and centralized control. For example, *beta testing* and *gated access* methods allow developers to make a model available only to a relatively small set of people before it is made fully open (Solaiman et al., 2025). Deploying models in stages can allow developers to gradually increase access while monitoring for unexpected uses and conducting research on potential harms. This allows a developer to refine their approach to safeguards and release before the model is fully open. This section focuses specifically on technical strategies that can serve as intermediate steps in staged deployments.

Existing work: There is a spectrum of deployment strategies between fully closed and fully open (Solaiman, 2023). Some of which, such as beta testing, do not have open technical problems related to open-weight

models and are thus out of scope of this paper (see Section 3).[4] However, here we consider several technical strategies on the openness spectrum.

First, *split deployment* strategies divide the model between client devices and server devices. Currently, most research on split learning and inference focuses on either enabling the use of large models on small devices (Xie et al., 2025; Lin et al., 2024; Ren et al., 2023) or keeping user inputs private from developers (Yao & Li, 2024; Mai et al., 2024; Shu et al., 2025).

Second, there more niche technical strategies for restricted forms of deployment that involve *hardware locking* (Clifford et al., 2025) or *homomorphic encryption* (Podschwadt et al., 2022).[5]

4.4.1. What exfiltration risks do split deployment strategies pose, and how can they be mitigated? Successful split deployments require that private model layers are kept secure. However, attackers can aim to exfiltrate them using reconstruction (e.g., Shu et al., 2025; Nevo et al., 2024; Carlini et al., 2024) or distillation (e.g., Huangpu & Gao, 2024) methods. Prior work on security focuses on securing client inputs against reconstruction attacks on small models Zhu et al. (2025b); Shu et al. (2025); Shabbir et al. (2025). In this regime, attackers have the upper hand – easily being able to reconstruct small, private portions of a model. In general, it is not well understood how the scale of the model and the proportion of parameters hidden change the cost of effectively reconstructing or distilling private layers of a model. Initial work on adapting this work to LLMs and frontier models broadly Shu et al. (2025); Yao & Li (2024) has also not measured how split location impacts vulnerability to attacks, nor tested the efficacy of a wide variety of attacks. It is not currently well understood how vulnerable private layers of a model are to exfiltration as a function of the architecture, hidden layers, model size, and attack algorithm.

4.4.2. How can we design split learning and inference APIs that are less costly and more competitive? A second challenge for split deployment strategies is the induced latency due to communication across the split. The necessity of overhead represents a fundamental limitation compared to other forms of deployment which can make split strategies less competitive. Efficiency challenges are especially acute for autoregressive and diffusion models, which require communication between the server and client for every iteration. This invites future work to develop competitive alternative models and methods Sahoo et al. (2024); Xie et al. (2025); Shen et al. (2025) that reduce the amount of information shared between the server and client, the number of messages passed, and/or the delay from overhead.

4.4.3. Can hardware locking or homomorphic encryption offer practical options for staged deployment? First, *hardware locking*, involves linking a model to specific, secure hardware (Clifford et al., 2025). This process certifies a model as "runnable" on a given piece of hardware, creating a secure chain of trust from the hardware to the model itself. Hardware-locking is precedented in traditional software, where hardware security enforcement is used to operate a zero-trust environment. However, designing and deploying infrastructure in an ever-changing open-weight model ecosystem would be challenging. The requirement for highly specialized and secure hardware poses a significant barrier to practical usage, and may be prohibitive for less-resourced developers. It is currently unclear if and how hardware locking could be a useful strategy for staged deployments. Meanwhile, at best, it could only be helpful for safety in niche applications. A second partially open strategy could involve a developer releasing a cryptographically encrypted model and retaining the exclusive ability to homomorphically encrypt inputs for it via an API. Current uses of these techniques focus on privacy-preserving machine learning rather than open-weight deployment (Lee et al., 2022; Podschwadt et al., 2022; Brand & Pradel, 2023; Ebel et al., 2025; Cheng et al., 2025b). However, scale poses a key practical challenge. Existing frameworks can only practically handle models with tens of millions

[4]Another such strategy is to use fine-tuning APIs which allow for users to experiment with fine-tuning a closed-weight model (Wu et al., 2024). There are open technical problems related to the safety of fine-tuning APIs such as reliably detecting adversarial attempts at obfuscating harmful fine-tuning data (Halawi et al., 2024; Davies et al., 2025) (see also Section 4.2 and Section 4.3). However, when using a fine-tuning API as a step for staged deployment, a model developer will typically not want to restrict fine-tuning data in order to monitor more realistic misuses of the model when it is deployed with open weights. Thus, malicious users would have little to no incentive to obfuscating harmful fine-tuning data. As such, these challenges are out of scope for this paper (see Section 3).

[5]Deployments involving split inference, hardware locking, or homomorphic encryption do not constitute "open-weight" releases in the traditional sense, as users cannot independently run the full model. We base our discussion of these strategies in this section on the premise that they can serve as intermediate steps in staged deployments that enable monitoring and risk assessment before a model's full open release.

of parameters (e.g., Ebel et al., 2025). It is not clear if homomorphic encryption offers a practical option for frontier models. Like hardware locking, it could only be useful in niche applications.

4.5 Model Provenance and Forensics

Model provenance methods help stakeholders study the spread and uses of open-weight models. While not directly upstream of model releases, ecosystem monitoring methods are a key component of risk management because they help stakeholders better study the real-world uses and impacts of models. Model provenance and forensics in the open-weight AI ecosystem are key to answering questions such as "What model is this?" and "What modifications has it undergone since its original release?"

Existing work: Here, we discuss three complementary types of methods: model watermarking, model heritage inference, and proof of training.

First, *model watermarking* methods aid in the identification of models. In contrast to data watermarking methods (Zhao et al., 2025b), model watermarks refer to model properties that serve to uniquely identify a model or a single instance of a model. Some approaches for model watermarking embed signals during generation without modifying the model's weights, but they depend on specialized decoding algorithms that can be disabled by users (Kirchenbauer et al., 2024a;b). Less-tamperable model watermarking methods must be 'baked into' a model's weights. For example, some methods allow for detection by implanting unique model *behaviors* (e.g., Yu et al., 2021; Fernandez et al., 2023; Xu et al., 2024; Christ et al., 2024). Other model watermarks allow for detection by analyzing model *parameters* by adding noise signatures across the model (e.g., Pagnotta et al., 2024; Block et al., 2025). Additional approaches have also been developed for quantization-based schemes (Li et al., 2023). However, surveys have emphasized ongoing challenges in balancing robustness and imperceptibility tradeoffs (Liang et al., 2024; Boenisch, 2021; Liu et al., 2024a).

Second, in contrast to watermarks, *model heritage inference* methods help researchers study the spread of models in the wild. These techniques have been used to reconstruct a genealogy from weights alone (Horwitz et al., 2025b). Zhu et al. (2025a) and Nikolic et al. (2025) also developed statistical tests determine if two models were trained independently or not. These techniques could offer useful tools to study real-world impacts and enforce licenses.

Finally, *proof of training* methods can be used for verifying that AI systems have undergone training processes with specific properties. Due to the complex supply chains behind some open-weight models, proof of training methods can uniquely enable trust throughout a model supply chain (Jia et al., 2021). Here, recent literature suggested ways to enable provable yet private training provenance using cryptographic methods to create verifiable, records of a model's origins (Garg et al., 2023; Abbaszadeh et al., 2024; Meiklejohn et al., 2025). These methods allow a model developer to verify that their model was trained according to a specific process. For example, they could prove they excluded a certain type of harmful data from their training corpus or applied a specific safety fine-tuning algorithm.

4.5.1. How can we watermark models in ways that are more durable against common modifications without side effects? Evaluations suggest that current content attribution watermarks can become undetectable under common open-weight model modifications, such as quantization, fine-tuning, model merging, and pruning (Gloaguen et al., 2025). There is also an absence of standardized benchmarks for comparing watermark durability across realistic combinations of model modifications, such as quantization followed by fine-tuning or merging followed by distillation (Li et al., 2023; Lv et al., 2023). Improving durability requires addressing tensions between watermark subtlety, persistence, and robustness. While some techniques (e.g., Pagnotta et al., 2024) demonstrate significant robustness to removal techniques, they remain vulnerable to distillation or sophisticated tampering attacks combining multiple modification types (Christ et al., 2024). The community has yet to see content watermarks designed for deployments where white-box access for fine-tuning, distillation, model merging, and quantization in sequence are standard practice. This highlights the additional downstream challenge of incorporating these methods into usage frameworks that take their rate of false positives and false negatives into account.

4.5.2. What algorithms can enable scalable and versatile model heritage inference? Ecosystem-wide heritage inference is desirable (Horwitz et al., 2025a) but not tractable with current infrastructure and

methods. For example, using current methods (Horwitz et al., 2025b), charting models across a platform such as Hugging Face would require millions of pairwise comparisons between models. While independence between two specific models is computationally inexpensive (Zhu et al., 2025a), continuous ecosystem-wide monitoring must accommodate daily uploads of potentially thousands of new models. Current methods also face four critical limitations beyond computational scaling when simply comparing two models. First, mixed heritage models created through weight averaging, model merging, or 'model soups' remain unaddressed despite growing prevalence. Existing approaches have focused primarily on finetuning and single-parent lineages. Second, cross-architecture techniques to account for processes such as knowledge distillation are neglected. While Zhu et al. (2025a)'s unconstrained setting enables some comparisons through proxy models, systematic handling of diverse architectures demands architecture-agnostic methods. Concurrently with this work, (Kuditipudi et al., 2025) has introduced a black-box heritage inference method with the potential to address this challenge. Third, accurately quantifying the degree of contribution from multiple different parent models requires granular attribution methods. Fourth, adversarial scenarios where actors actively obscure provenance would require more robust methods. A final challenge will be the implementation of heritage inference in ways that are efficient and acccount for their false positives and negatives.

4.5.3. How practical and scalable are proof of training methods? Present techniques for proof of training have limitations (Choi et al., 2023; Sun et al., 2025). A major barrier to the practical adoption of provable provenance is computational overhead. Generating zero-knowledge proofs for training runs that involve trillions of datapoints and billions of model parameters is currently computationally prohibitive. The generated proofs must also be integrated into the broader AI ecosystem. This involves creating infrastructure for issuing, storing, and verifying cryptographic certificates. Finally, current research primarily focuses on verifying straightforward properties of the training process, such as the inclusion or exclusion of specific data. However, many advanced safety techniques involve nuanced procedures that are difficult to formalize and verify cryptographically. An open challenge is to extend proof of training methods to cover more qualitative safety-related interventions. However, at best, even if proof of training methods can be practically scaled and implemented, they could only be useful for safety in niche applications.

5 What techniques are prominent open-weight developers reporting on?

To understand what frontier open-weight model developers have reported about technical safeguards, we analyzed technical reports and model cards from popular open-weight models. We selected two sets of models. First, we identified the 10 most widely adopted open-weight models on Hugging Face. We selected these models by examining Hugging Face download statistics[6] as of Oct 15 2025: specifically, we selected the top 10 organizations by total model downloads (all time) that released foundation models[7] in 2025, then identified the most downloaded model from each organization. For model families released under shared documentation (e.g., Qwen3-0.6B, Qwen3-4B, Qwen3-8B), we report on the model family as a whole. This yielded the following models: Qwen3 (Yang et al., 2025), DeepSeek-R1 (Guo et al., 2025), Gemma3 (Team et al., 2025a), gpt-oss (Agarwal et al., 2025), Nemotron-Nano (Basant et al., 2025), Granite-3.3 (Granite Team, 2024), Phi-4 (Abdin et al., 2024), EXAONE-Deep (LG et al., 2025), Llada-8B (Nie et al., 2025), and GLM-4.5Team et al. (2025b).

Second, we examined specific image and video generation models that Hawkins et al. (2025) and Kamachee et al. (2025) highlighted as being commonly used for image and video deepfakes. These models included: Stable Diffusion 1.x (Rombach et al., 2021), FLUX (Batifol et al., 2025),[8] Wan2.x (Wan et al., 2025), HunyuanVideo (Kong et al., 2025), and LTXV (HaCohen et al., 2024).

While not exhaustive, these models represent both highly-adopted releases and models with documented misuse patterns, spanning multiple organizations, jurisdictions, architectures, and modalities.

[6] `https://huggingface.co/spaces/evijit/ModelVerse`

[7] We specifically look at models that are supported by the `text-generation` pipeline of the `transformer` library, as these constitute the vast majority of foundation models in popular use. Some of these models natively support multimodality.

[8] Batifol et al. (2025) postdates Hawkins et al. (2025), but Batifol et al. (2025) is the only technical report available for any FLUX model, so we analyze it in the Table 2.

Model	Organization	Safe Data Curation (Section 4.1)	Tamper-Resistance Training (Section 4.2)	Tampering Evals (Section 4.3)	Staged Deployment (Section 4.4)	Model Provenance (Section 4.5)
Most-Downloaded LLMs/Multimodal Foundation Models on Hugging Face that were released in 2025						
Qwen3 (Yang et al., 2025)	Alibaba	1-3 Sentences	No Mention	No Mention	No Mention	No Mention
DeepSeek-R1 (Guo et al., 2025)	DeepSeek	No Mention	No Mention	No Mention	No Mention	No Mention
Gemma3 (Team et al., 2025a)	Google	1-3 Sentences	No Mention	No Mention	No Mention	No Mention
gpt-oss (Agarwal et al., 2025)	OpenAI	Paragraph	No Mention	Dedicated Paper	No Mention	No Mention
Nemotron-Nano (Basant et al., 2025)	NVIDIA	Paragraph	No Mention	No Mention	No Mention	No Mention
Granite-3.3 (Granite Team, 2024)	IBM	Dedicated Section	No Mention	No Mention	No Mention	No Mention
Phi-4 (Abdin et al., 2024)	Microsoft	Paragraph	No Mention	No Mention	No Mention	No Mention
EXAONE-Deep (LG et al., 2025)	LG AI Research	No Mention	No Mention	No Mention	No Mention	No Mention
Llada-8B (Nie et al., 2025)	GSAI-ML	No Mention	No Mention	No Mention	No Mention	No Mention
GLM-4.5 (Team et al., 2025b)	Z.AI	No Mention	No Mention	No Mention	No Mention	No Mention
Models highlighted in Hawkins et al. (2025) and Kamachee et al. (2025)						
Stable Diffusion 1.x (Rombach et al., 2021)	Stability AI	No Mention	No Mention	No Mention	No Mention	No Mention
FLUX (Batifol et al., 2025)	Black Forest Labs	1-3 Sentences	No Mention	No Mention	No Mention	No Mention
Wan2.x (Wan et al., 2025)	Alibaba	1-3 Sentences	No Mention	No Mention	No Mention	No Mention
Stable Video Diffusion (Blattmann et al., 2023)	Stability AI	No Mention	No Mention	No Mention	No Mention	No Mention
HunyuanVideo (Kong et al., 2025)	Tencent	No Mention	No Mention	No Mention	No Mention	No Mention
LTX-Video (HaCohen et al., 2024)	Lightricks	No Mention	No Mention	No Mention	No Mention	No Mention

Table 2: **What technical safety techniques are prominent open-weight model developers reporting on?** We overview what open-weight model risk management techniques are discussed in technical reports. The top section includes the top 10 most-downloaded language/multimodal model families released between January and October 2025 (by organization). The bottom section includes image and video generation models highlighted by Hawkins et al. (2025) and Kamachee et al. (2025) as being prominently used for image and video deepfakes. This table offers a source of reference documenting developer disclosures. This table does not analyze substance and is not intended to be a scorecard. *Legend:* NM No Mention, 1-3S 1-3 Sentences, P Paragraph, DS Dedicated Section/Paper. "No mention" does not imply no implementation. We only focus on safety – for example, we do not analyze reporting on quality-focused data curation.

Summarizing reporting on technical safeguards: For each of the five categories of safeguards discussed in Section 3, we examined whether each model's documentation reported on the use of these techniques for improving safety. We categorized reporting as: **no mention**, a **1-3 sentence mention**, a **paragraph-level description**, or a **dedicated section/paper**. We note that this table is qualitative in nature and is meant to analyze overall safety reporting trends by open model developers without necessarily highlighting any particular organization or model. "No mention" does not imply "not implemented", and our analysis does not consider the substance or effectiveness of reported techniques. Our observations are shown in Table 2.

Our analysis reveals several patterns in how open-weight model developers report on technical safeguards. Among the top open model developers by downloads of models released in 2025, data curation is the most commonly reported safeguard, with 6 out of 10 models providing at least brief documentation—ranging from 1-3 sentences (Qwen3, Gemma3) to dedicated sections (Granite-3.3). Three models provide paragraph-level descriptions (gpt-oss, Nemotron-Nano, Phi-4), though notably some of these focus on post-training or mid-training safety fine-tuning rather than pre-training data filtering specifically which may be key for tamper-resistent safeguards (see Section 4.2). However, documentation for other technical open-weight

model safeguards remains sparse. Tamper-resistant training algorithms recieve no mention in any of the analyzed models. Tampering evaluations appear in only one model (gpt-oss), which dedicated a separate paper to adversarial fine-tuning assessments (Wallace et al., 2025). Staged deployment strategies and model provenance/forensics techniques are absent from all technical reports examined.[9]

Among the five image and video generation models highlighted in Hawkins et al. (2025) and (Kamachee et al., 2025), similarly few safeguards are reported across models with only the FLUX and Wan2.x technical reports making mention of safety-focused data curation (Batifol et al., 2025; Wan et al., 2025).

Implications: The prevalence of grey cells in Table 2 suggests substantial room for growth in the science of technical open-weight model safety. The general absence of reporting on tamper-resistance, tampering evaluations, and provenance techniques is particularly notable given the vulnerability of open-weight models to tampering attacks. This gap between documented risks and reported mitigations suggests that either: (1) these techniques are not being widely implemented, (2) they are being implemented but not documented, or (3) effective methods for these safeguards remain underdeveloped.

These findings align with our broader argument that building the science of open-weight model risk management requires not only developing new technical safeguards, but also establishing norms around transparent reporting of safety practices. The scarcity of documentation across multiple safeguard categories suggests the field could benefit from more thorough reporting on technical risk mitigation strategies.

6 Discussion

Significance: Increasingly capable open-weight models are being released on a regular basis, with research showing open-weight model capabilities to consistently be only 6-12 months behind frontier proprietary models (Cottier, 2024; Maslej et al., 2025). There are clear benefits of open-weight model development. These include driving innovation, enabling AI safety/security research, enabling flexible AI adoption, and spreading benefits and access to AI (Seger & Hancock, 2025). However, open-weight models also pose distinct risks stemming from the potential for rapid proliferation of model flaws and the ease with which malicious actors can bypass safeguards against misuse. We believe that a positive future with AI will involve a balance of proprietary and open-weight model development. Effective tools to mitigate risks will not only be key for mitigating open-weight models' risks but also accessing their benefits by avoiding backlash (Henderson et al., 2023). Toward this end, this paper investigates technical interventions that could help mitigate and monitor risks from open-weight AI models. Our collective hope is that this paper will help to build the field of technical open-weight model risk management.

Limitations: As discussed in Section 3, this paper only focuses on technical tools with distinct implications for open-weight models. This focus is not meant to imply that open problems, strategies distinct to open-weight models, or technical strategies are the most useful or important. We concur with François et al. (2025), Srikumar et al. (2024), and UK AISI (2025) that a holistic approach to monitoring and mitigating risks in the open-weight model ecosystem will be crucial. However, not all techniques for open-weight model safety will be equally effective or competitive. Research on the limitations and practicality of techniques will be important for refining the toolkit.

Uncertainties: It is unclear how effective different safeguards for open-weight models will ultimately be. Not all approaches will be equally effective. It is also unclear how much counterfactual risk open-weight models will pose compared to closed-weight models (Kapoor et al., 2024). Thus, we emphasize the value of gathering more information about open-weight models through additional research and analysis of impacts across the ecosystem. In doing so, the research community should be mindful of both 'openness washing' (Grieve, 2024) and 'safety washing' (Ren et al., 2024). It is important for researchers and policymakers to be open to evidence both in favor of and against the possibility that some models may pose large risks if deployed in certain ways. Some models – even with safeguards – might enable acute misuse if deployed

[9]Several organizations in Table 2 have implemented safeguards not specific to open-weight models and/or released companion safety/guardrail models alongside their base models, including Qwen3Guard Zhao et al. (2025a), Granite Guardian Padhi et al. (2024), and NemoGuard Rebedea et al. (2023). While these external safety tools are out of scope for this analysis, they represent meaningful contributions to the open-weight safety ecosystem.

with open weights. Others might significantly hinder open-science or concentrate large amounts of power if deployed with closed weights. Others still might pose major risks regardless of deployment type.

Incentivizing future research: While we are optimistic about the potential value of more research into technical mitigations against open-weight model risks, we recognize that incentives for private actors to research and develop robust safeguards for frontier open-weight models are currently limited. Furthermore, technical safeguards for open models will only be resistant to some degree of intervention. So from a researcher's standpoint, work on technical interventions may be high-risk (in terms of investment) and limited reward. This does not mean, however, that this work is not worthwhile. Each of the strategies we discussed in Section 4 is individually imperfect, but contributes meaningfully to reducing harm or increasing information. Used in concert, these methods can substantially improve risk management. There are also barriers to important safety and security research that remain in place. While many open developers provide models with the intent of enabling researcher, and participate in open flaw bounties (McGregor et al., 2025), some major open-weight developers do not consistently offer legal 'safe harbors' and even impose legal language or technical obstacles against good-faith safety evaluations into their systems' safeguards (Longpre et al., 2024a).

The importance of openness (not just of model weights): The status quo may currently incentivize little openness related to open-weight model risk management (Table 2). However, in building the science of open-weight model risk management, we emphasize the value of open scientific collaboration (Phang et al., 2022; Linåker et al., 2025; Scotti, 2025), open research (Biderman et al., 2023; Liu et al., 2023; Groeneveld et al., 2024), open evaluations (Gao et al., 2021; Bommasani et al., 2023; Biderman et al., 2024), open reporting about risk-management methodology (Seger, 2024), and open standardized documentation. Just as building the science of open-weight model risk management will provide a collective good, it will also require collective effort.

Acknowledgments

We are thankful to Anka Reuel, Isabella Duan, Jack Sanderson, Nicholas Carlini, and Stella Biderman for discussions on drafts of the paper.

References

Kasra Abbaszadeh, Christodoulos Pappas, Jonathan Katz, and Dimitrios Papadopoulos. Zero-knowledge proofs of training for deep neural networks. Cryptology ePrint Archive, Paper 2024/162, 2024. URL `https://eprint.iacr.org/2024/162`.

Amro Abdalla, Ismail Shaheen, Dan DeGenaro, Rupayan Mallick, Bogdan Raita, and Sarah Adel Bargal. GIFT: Gradient-aware Immunization of diffusion models against malicious Fine-Tuning with safe concepts retention, July 2025. URL `http://arxiv.org/abs/2507.13598`. arXiv:2507.13598 [cs].

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

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

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. A survey on data selection for language models. *Transactions on Machine Learning Research*.

Rico Angell, Jannik Brinkmann, and He He. Jailbreak Strength and Model Similarity Predict Transferability, June 2025. URL `http://arxiv.org/abs/2506.12913`. arXiv:2506.12913 [cs].

Anonymous. Antibody: Strengthening defense against harmful fine-tuning for large language models via attenuating harmful gradient influence. In *Submitted to The Fourteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=qur2ef8MqQ. under review.

Anonymous. Gradshield: Alignment preserving finetuning. In *Submitted to The Fourteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=YYUNm7IibC. under review.

Anthropic. Activating ai safety level 3 protections. Technical report, Anthropic, May 2025. URL https://www-cdn.anthropic.com/807c59454757214bfd37592d6e048079cd7a7728.pdf.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*, 2024.

Yang Bai, Gaojie Xing, Hongyan Wu, Zhihong Rao, Chuan Ma, Shiping Wang, Xiaolei Liu, Yimin Zhou, Jiajia Tang, Kaijun Huang, et al. Backdoor attack and defense on deep learning: A survey. *IEEE Transactions on Computational Social Systems*, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, April 2022. URL http://arxiv.org/abs/2204.05862. arXiv:2204.05862 [cs].

Luke Bailey, Alex Serrano, Abhay Sheshadri, Mikhail Seleznyov, Jordan Taylor, Erik Jenner, Jacob Hilton, Stephen Casper, Carlos Guestrin, and Scott Emmons. Obfuscated Activations Bypass LLM Latent-Space Defenses, February 2025. URL http://arxiv.org/abs/2412.09565. arXiv:2412.09565 [cs].

Fazl Barez, Tingchen Fu, Ameya Prabhu, Stephen Casper, Amartya Sanyal, Adel Bibi, Aidan O'Gara, Robert Kirk, Ben Bucknall, Tim Fist, Luke Ong, Philip Torr, Kwok-Yan Lam, Robert Trager, David Krueger, Sören Mindermann, José Hernandez-Orallo, Mor Geva, and Yarin Gal. Open Problems in Machine Unlearning for AI Safety, January 2025. URL http://arxiv.org/abs/2501.04952. arXiv:2501.04952 [cs].

Aarti Basant, Abhijit Khairnar, Abhijit Paithankar, Abhinav Khattar, Adithya Renduchintala, Aditya Malte, Akhiad Bercovich, Akshay Hazare, Alejandra Rico, Aleksander Ficek, et al. Nvidia nemotron nano 2: An accurate and efficient hybrid mamba-transformer reasoning model. *arXiv preprint arXiv:2508.14444*, 2025.

Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pp. arXiv–2506, 2025.

Yoshua Bengio, Tegan Maharaj, Luke Ong, Stuart Russell, Dawn Song, Max Tegmark, Lan Xue, Ya-Qin Zhang, Stephen Casper, Wan Sie Lee, Sören Mindermann, et al. The Singapore Consensus on Global AI Safety Research Priorities, June 2025a. URL http://arxiv.org/abs/2506.20702. arXiv:2506.20702 [cs].

Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, et al. International ai safety report. *arXiv preprint arXiv:2501.17805*, 2025b.

Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. Taken out of context: On measuring situational awareness in LLMs, September 2023. URL http://arxiv.org/abs/2309.00667. arXiv:2309.00667 [cs].

Kushal Raj Bhandari, Pin-Yu Chen, and Jianxi Gao. Forecasting open-weight ai model growth on huggingface. *arXiv preprint arXiv:2502.15987*, 2025.

Rishabh Bhardwaj and Soujanya Poria. Language Model Unalignment: Parametric Red-Teaming to Expose Hidden Harms and Biases, November 2023. URL http://arxiv.org/abs/2310.14303. arXiv:2310.14303 [cs].

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, et al. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Adam Block, Ayush Sekhari, and Alexander Rakhlin. GaussMark: A Practical Approach for Structural Watermarking of Language Models, January 2025. URL `http://arxiv.org/abs/2501.13941`. arXiv:2501.13941 [cs].

Franziska Boenisch. A Systematic Review on Model Watermarking for Neural Networks. *Frontiers in Big Data*, 4:729663, November 2021. ISSN 2624-909X. doi: 10.3389/fdata.2021.729663. URL `http://arxiv.org/abs/2009.12153`. arXiv:2009.12153 [cs].

Rishi Bommasani, Percy Liang, and Tony Lee. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146, 2023.

Rishi Bommasani, Sayash Kapoor, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Daniel Zhang, Marietje Schaake, Daniel E. Ho, Arvind Narayanan, and Percy Liang. Considerations for governing open foundation models. *Science*, 386(6718):151–153, October 2024. doi: 10.1126/science.adp1848. URL `https://www.science.org/doi/abs/10.1126/science.adp1848`. Publisher: American Association for the Advancement of Science.

Michael Brand and Gaëtan Pradel. Practical privacy-preserving machine learning using fully homomorphic encryption. *Cryptology ePrint Archive*, 2023.

Davis Brown, Mahdi Sabbaghi, Luze Sun, Alexander Robey, George J. Pappas, Eric Wong, and Hamed Hassani. Benchmarking Misuse Mitigation Against Covert Adversaries, June 2025. URL `http://arxiv.org/abs/2506.06414`. arXiv:2506.06414 [cs].

Wenjun Cao. Fight fire with fire: Defending against malicious rl fine-tuning via reward neutralization. *arXiv preprint arXiv:2505.04578*, 2025.

Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, et al. Stealing part of a production language model. *arXiv preprint arXiv:2403.06634*, 2024.

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, et al. Black-Box Access is Insufficient for Rigorous AI Audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 2254–2272, New York, NY, USA, June 2024. Association for Computing Machinery. ISBN 979-8-4007-0450-5. doi: 10.1145/3630106.3659037. URL `https://dl.acm.org/doi/10.1145/3630106.3659037`.

Sven Cattell, Avijit Ghosh, and Lucie-Aimée Kaffee. Coordinated Flaw Disclosure for AI: Beyond Security Vulnerabilities. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):267–280, October 2024. ISSN 3065-8365. doi: 10.1609/aies.v7i1.31635. URL `https://ojs.aaai.org/index.php/AIES/article/view/31635`.

Sarah H. Cen, Aspen Hopkins, Andrew Ilyas, Aleksander Madry, Isabella Struckman, and Luis Videgaray. Ai supply chains (and why they matter), April 2023. URL `https://aipolicy.substack.com/p/supply-chains-2`. The second post in our series On AI Deployment.

Alan Chan, Ben Bucknall, Herbie Bradley, and David Krueger. Hazards from Increasingly Accessible Fine-Tuning of Downloadable Foundation Models, December 2023. URL http://arxiv.org/abs/2312.14751.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. How Do Large Language Models Acquire Factual Knowledge During Pretraining?, November 2024. URL http://arxiv.org/abs/2406.11813. arXiv:2406.11813 [cs].

Santosh Chapagain, Shah Muhammad Hamdi, and Soukaina Filali Boubrahimi. Pruning strategies for backdoor defense in llms. *arXiv preprint arXiv:2508.20032*, 2025.

Zora Che, Stephen Casper, Robert Kirk, Anirudh Satheesh, Stewart Slocum, Lev E. McKinney, Rohit Gandikota, Aidan Ewart, Domenic Rosati, Zichu Wu, Zikui Cai, Bilal Chughtai, Yarin Gal, Furong Huang, and Dylan Hadfield-Menell. Model Tampering Attacks Enable More Rigorous Evaluations of LLM Capabilities, July 2025. URL http://arxiv.org/abs/2502.05209. arXiv:2502.05209 [cs].

Kejia Chen, Jiawen Zhang, Jiacong Hu, Yu Wang, Jian Lou, Zunlei Feng, and Mingli Song. Q-resafe: Assessing Safety Risks and Quantization-aware Safety Patching for Quantized Large Language Models, June 2025a. URL http://arxiv.org/abs/2506.20251. arXiv:2506.20251 [cs].

Liang Chen, Xueting Han, Li Shen, Jing Bai, and Kam-Fai Wong. Vulnerability-aware alignment: Mitigating uneven forgetting in harmful fine-tuning. *arXiv preprint arXiv:2506.03850*, 2025b.

Yanda Chen, Mycal Tucker, Nina Panickssery, Tony Wang, Francesco Mosconi, Anjali Gopal, Carson Denison, Linda Petrini, Jan Leike, Ethan Perez, and Mrinank Sharma. Enhancing model safety through pretraining data filtering. *Anthropic Alignment Blog*, aug 2025c. URL https://alignment.anthropic.com/2025/pretraining-data-filtering/.

Zehua Cheng, Manying Zhang, Jiahao Sun, and Wei Dai. On weaponization-resistant large language models with prospect theoretic alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 10309–10324, 2025a.

Zerui Cheng, Edoardo Contente, Oleg Aleksandrovich Golev, Jonathan Hayase, Andrew Miller, Niusha Moshrefi, Anshul Nasery, Sewoong Oh, Himanshu Tyagi, and Pramod Viswanath. Position: Can openai be truly" open"? open, monetizable, and loyal (oml) ai is what you need in a fair and sustainable next-generation ai ecosystem. 2025b.

Dami Choi, Yonadav Shavit, and David K Duvenaud. Tools for verifying neural models' training data. *Advances in Neural Information Processing Systems*, 36:1154–1188, 2023.

Miranda Christ, Sam Gunn, Tal Malkin, and Mariana Raykova. Provably Robust Watermarks for Open-Source Language Models, October 2024. URL http://arxiv.org/abs/2410.18861. arXiv:2410.18861 [cs].

Eleanor Clifford, Adhithya Saravanan, Harry Langford, Cheng Zhang, Yiren Zhao, Robert Mullins, Ilia Shumailov, and Jamie Hayes. Locking Machine Learning Models into Hardware. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 302–320. IEEE Computer Society, April 2025. ISBN 979-8-3315-1711-3. doi: 10.1109/SaTML64287.2025.00023. URL https://www.computer.org/csdl/proceedings-article/satml/2025/171100a302/26VnsgjHTZ6.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the Ripple Effects of Knowledge Editing in Language Models. *Transactions of the Association for Computational Linguistics*, 12:283–298, 2024. doi: 10.1162/tacl_a_00644. URL https://aclanthology.org/2024.tacl-1.16/. Place: Cambridge, MA Publisher: MIT Press.

A Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. Accountability in an algorithmic society: relationality, responsibility, and robustness in machine learning. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 864–876, 2022.

Ben Cottier. How far behind are open models?, November 2024. URL https://epoch.ai/blog/open-models-report.

Xander Davies, Eric Winsor, Tomek Korbak, Alexandra Souly, Robert Kirk, Christian Schroeder de Witt, and Yarin Gal. Fundamental Limitations in Defending LLM Finetuning APIs, February 2025. URL http://arxiv.org/abs/2502.14828. arXiv:2502.14828 [cs].

Aghyad Deeb and Fabien Roger. Do Unlearning Methods Remove Information from Language Model Weights?, February 2025. URL http://arxiv.org/abs/2410.08827. arXiv:2410.08827 [cs].

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus, September 2021. URL http://arxiv.org/abs/2104.08758. arXiv:2104.08758 [cs].

Ann-Kathrin Dombrowski, Dillon Bowen, Adam Gleave, and Chris Cundy. The Safety Gap Toolkit: Evaluating Hidden Dangers of Open-Source Models, July 2025. URL http://arxiv.org/abs/2507.11544. arXiv:2507.11544 [cs].

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zachary C Lipton, J Zico Kolter, and Pratyush Maini. Openunlearning: Accelerating llm unlearning via unified benchmarking of methods and metrics. *arXiv preprint arXiv:2506.12618*, 2025.

Austin Ebel, Karthik Garimella, and Brandon Reagen. Orion: A fully homomorphic encryption framework for deep learning. In *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pp. 734–749, 2025.

Kazuki Egashira, Mark Vero, Robin Staab, Jingxuan He, and Martin Vechev. Exploiting LLM Quantization, November 2024. URL http://arxiv.org/abs/2405.18137. arXiv:2405.18137 [cs].

Francisco Eiras, Aleksandar Petrov, Philip HS Torr, M Pawan Kumar, and Adel Bibi. Do as i do (safely): Mitigating task-specific fine-tuning risks in large language models. *arXiv preprint arXiv:2406.10288*, 2024a.

Francisco Eiras, Aleksandar Petrov, Bertie Vidgen, Christian Schroeder, Fabio Pizzati, Katherine Elkins, Supratik Mukhopadhyay, Adel Bibi, et al. Risks and Opportunities of Open-Source Generative AI, May 2024b. URL http://arxiv.org/abs/2405.08597. arXiv:2405.08597 [cs].

Chongyu Fan, Jinghan Jia, Yihua Zhang, Anil Ramakrishna, Mingyi Hong, and Sijia Liu. Towards LLM Unlearning Resilient to Relearning Attacks: A Sharpness-Aware Minimization Perspective and Beyond, May 2025a. URL http://arxiv.org/abs/2502.05374. arXiv:2502.05374 [cs].

Chongyu Fan, Changsheng Wang, Yancheng Huang, Soumyadeep Pal, and Sijia Liu. Llm unlearning under the microscope: A full-stack view on methods and metrics. *arXiv preprint arXiv:2510.07626*, 2025b.

Weitao Feng, Lixu Wang, Tianyi Wei, Jie Zhang, Chongyang Gao, Sinong Zhan, Peizhuo Lv, and Wei Dong. Token buncher: Shielding llms from harmful reinforcement learning fine-tuning. *arXiv preprint arXiv:2508.20697*, 2025.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The Stable Signature: Rooting Watermarks in Latent Diffusion Models, July 2023. URL http://arxiv.org/abs/2303.15435. arXiv:2303.15435 [cs].

Camille François, Ludovic Péran, Ayah Bdeir, Nouha Dziri, Will Hawkins, Yacine Jernite, Sayash Kapoor, Juliet Shen, Heidy Khlaaf, Kevin Klyman, et al. A Different Approach to AI Safety: Proceedings from the Columbia Convening on Openness in Artificial Intelligence and AI Safety, June 2025. URL http://arxiv.org/abs/2506.22183. arXiv:2506.22183 [cs].

Yarin Gal. Towards a science of ai evaluations. Blog post, 2024. URL https://www.cs.ox.ac.uk/people/yarin.gal/website/blog_98A8.html.

Yarin Gal and Stephen Casper. Customizable ai systems that anyone can adapt bring big opportunities—and even bigger risks. *Nature*, 646(8084):286–287, 2025.

Hongcheng Gao, Tianyu Pang, Chao Du, Taihang Hu, Zhijie Deng, and Min Lin. Meta-unlearning on diffusion models: Preventing relearning unlearned concepts. *arXiv preprint arXiv:2410.12777*, 2024.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Leo Gao, Jonathan Tow, Stella Biderman, Shawn Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jasmine Hsu, Kyle McDonell, Niklas Muennighoff, et al. A framework for few-shot language model evaluation. *Version v0. 0.1. Sept*, 10:8–9, 2021.

Sanjam Garg, Aarushi Goel, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, Guru-Vamsi Policharla, and Mingyuan Wang. Experimenting with zero-knowledge proofs of training. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, CCS '23, pp. 1880–1894, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400700507. doi: 10.1145/3576915.3623202. URL https://doi.org/10.1145/3576915.3623202.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting Recall of Factual Associations in Auto-Regressive Language Models, October 2023. URL http://arxiv.org/abs/2304.14767. arXiv:2304.14767 [cs].

Thibaud Gloaguen, Nikola Jovanović, Robin Staab, and Martin Vechev. Towards Watermarking of Open-Source LLMs, February 2025. URL http://arxiv.org/abs/2502.10525. arXiv:2502.10525 [cs].

Goodfire. Under the hood of a reasoning model. April 2025. URL https://www.goodfire.ai/research/under-the-hood-of-a-reasoning-model.

Google AI. Gemini 2.5 pro preview model card. Model card / technical report, Google, May 2025.

IBM Granite Team. Granite 3.0 language models. *URL: https://github. com/ibm-granite/granite-3.0-language-models*, 2024.

Ryan Greenblatt, Fabien Roger, Dmitrii Krasheninnikov, and David Krueger. Stress-Testing Capability Elicitation With Password-Locked Models, May 2024. URL http://arxiv.org/abs/2405.19550. arXiv:2405.19550 [cs].

Elisabeth Grieve. Openness hype and open washing: A critical analysis of openness discourses in generative ai. *McMaster University Theses*, 2024.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamilė Lukošiūtė, et al. Studying Large Language Model Generalization with Influence Functions, August 2023. URL http://arxiv.org/abs/2308.03296. arXiv:2308.03296 [cs].

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z. Publisher: Nature Publishing Group.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Danny Halawi, Alexander Wei, Eric Wallace, Tony T. Wang, Nika Haghtalab, and Jacob Steinhardt. Covert Malicious Finetuning: Challenges in Safeguarding LLM Adaptation, June 2024. URL http://arxiv.org/abs/2406.20053. arXiv:2406.20053 [cs].

Seokil Ham, Yubin Choi, Yujin Yang, Seungju Cho, Younghun Kim, and Changick Kim. Safety-aligned weights are not enough: Refusal-teacher-guided finetuning enhances safety and downstream performance under harmful finetuning attacks. 2025. URL https://arxiv.org/abs/2506.07356.

Kimia Hamidieh, Lester Mackey, and David Alvarez-Melis. Domain-aware scaling laws uncover data synergy. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025.

Hasan Abed Al Kader Hammoud, Umberto Michieli, Fabio Pizzati, Philip Torr, Adel Bibi, Bernard Ghanem, and Mete Ozay. Model Merging and Safety Alignment: One Bad Model Spoils the Bunch, June 2024. URL http://arxiv.org/abs/2406.14563. arXiv:2406.14563 [cs].

Muhammad Abdullah Hanif, Nandish Chattopadhyay, Bassem Ouni, and Muhammad Shafique. Survey on backdoor attacks on deep learning: Current trends, categorization, applications, research challenges, and future prospects. *IEEE Access*, 2025.

Peter Hase, Thomas Hofweber, Xiang Zhou, Elias Stengel-Eskin, and Mohit Bansal. Fundamental Problems With Model Editing: How Should Rational Belief Revision Work in LLMs?, June 2024. URL http://arxiv.org/abs/2406.19354. arXiv:2406.19354 [cs].

Will Hawkins, Chris Russell, and Brent Mittelstadt. Deepfakes on Demand: the rise of accessible non-consensual deepfake image generators, May 2025. URL http://arxiv.org/abs/2505.03859. arXiv:2505.03859 [cs].

Luxi He, Mengzhou Xia, and Peter Henderson. What is in your safe data? identifying benign data that breaks safety. *arXiv preprint arXiv:2404.01099*, 2024.

Yichen He, Guanhua Huang, Peiyuan Feng, Yuan Lin, Yuchen Zhang, Hang Li, and Weinan E. PaSa: An LLM Agent for Comprehensive Academic Paper Search, May 2025. URL http://arxiv.org/abs/2501.10120. arXiv:2501.10120 [cs].

Peter Henderson, Eric Mitchell, Christopher D. Manning, Dan Jurafsky, and Chelsea Finn. Self-Destructing Models: Increasing the Costs of Harmful Dual Uses of Foundation Models, August 2023. URL http://arxiv.org/abs/2211.14946. arXiv:2211.14946 [cs].

Felix Hofstätter, Teun van der Weij, Jayden Teoh, Rada Djoneva, Henning Bartsch, and Francis Rhys Ward. The Elicitation Game: Evaluating Capability Elicitation Techniques, July 2025. URL http://arxiv.org/abs/2502.02180. arXiv:2502.02180 [cs].

Eliahu Horwitz, Nitzan Kurer, Jonathan Kahana, Liel Amar, and Yedid Hoshen. We should chart an atlas of all the world's models. *arXiv preprint arXiv:2503.10633*, 2025a.

Eliahu Horwitz, Asaf Shul, and Yedid Hoshen. Unsupervised Model Tree Heritage Recovery, April 2025b. URL http://arxiv.org/abs/2405.18432. arXiv:2405.18432 [cs].

Saad Hossain, Samanvay Vajpayee, and Sirisha Rambhatla. SafeTuneBed: A Toolkit for Benchmarking LLM Safety Alignment in Fine-Tuning, May 2025. URL http://arxiv.org/abs/2506.00676. arXiv:2506.00676 [cs].

Lei Hsiung, Tianyu Pang, Yung-Chen Tang, Linyue Song, Tsung-Yi Ho, Pin-Yu Chen, and Yaoqing Yang. Why llm safety guardrails collapse after fine-tuning: A similarity analysis between alignment and fine-tuning datasets. *arXiv preprint arXiv:2506.05346*, 2025.

Nathan Hu, Benjamin Wright, Carson Denison, Sam Marks, Johannes Treutlein, Jonathan Uesato, and Evan Hubinger. Training on documents about reward hacking induces reward hacking. Anthropic Alignment Science Blog, 2025a. URL https://alignment.anthropic.com/2025/reward-hacking-ooc/. Accessed: 2025-10-07.

Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith. Unlearning or Obfuscating? Jogging the Memory of Unlearned LLMs via Benign Relearning, March 2025b. URL http://arxiv.org/abs/2406.13356. arXiv:2406.13356 [cs].

Zixuan Hu, Li Shen, Zhenyi Wang, Yongxian Wei, and Dacheng Tao. Adaptive defense against harmful fine-tuning for large language models via bayesian data scheduler. *arXiv preprint arXiv:2510.27172*, 2025c.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Booster: Tackling harmful fine-tuning for large language models via attenuating harmful perturbation. *arXiv preprint arXiv:2409.01586*, 2024a.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Harmful Fine-tuning Attacks and Defenses for Large Language Models: A Survey, December 2024b. URL http://arxiv.org/abs/2409.18169.

Tiansheng Huang, Sihao Hu, and Ling Liu. Vaccine: Perturbation-aware alignment for large language models against harmful fine-tuning attack. *Advances in Neural Information Processing Systems*, 37:74058–74088, 2024c.

Qionglin Huangpu and Huixiang Gao. Efficient model compression and knowledge distillation on llama 2: Achieving high performance with reduced computational cost. 2024.

IWF. What has changed in the ai csam landscape?
ai csam report update, july 2024. Technical report, Internet Watch Foundation, July 2024. URL https://admin.iwf.org.uk/media/nadlcb1z/iwf-ai-csam-report_update-public-jul24v13.pdf.

Mona Jami Pour, Seyed Mohammadbagher Jafari, and Monireh Khani. How to Know Your Customers? Towards a Novel Framework for Online Customer Knowledge Absorptive Capacity. *Journal of the Knowledge Economy*, December 2024. ISSN 1868-7873. doi: 10.1007/s13132-024-02533-4. URL https://doi.org/10.1007/s13132-024-02533-4.

Hengrui Jia, Mohammad Yaghini, Christopher A. Choquette-Choo, Natalie Dullerud, Anvith Thudi, Varun Chandrasekaran, and Nicolas Papernot. Proof-of-learning: Definitions and practice. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 1039–1056, 2021. doi: 10.1109/SP40001.2021.00106.

Max Kamachee, Stephen Casper, Michelle L. Ding, Rui-Jie Yew, Anka Reuel, Stella Biderman, and Dylan Hadfield-Menell. Video deepfake abuse: How company choices predictably shape misuse patterns. *Available at SSRN*, November 2025. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5829303. SSRN ID: 5829303.

Nikhil Kandpal, Brian Lester, Colin Raffel, Sebastian Majstorovic, Stella Biderman, Baber Abbasi, Luca Soldaini, Enrico Shippole, A Feder Cooper, Aviya Skowron, et al. The common pile v0. 1: An 8tb dataset of public domain and openly licensed text. *arXiv preprint arXiv:2506.05209*, 2025.

Sayash Kapoor, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, et al. On the Societal Impact of Open Foundation Models, February 2024. URL http://arxiv.org/abs/2403.07918. arXiv:2403.07918 [cs].

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A Watermark for Large Language Models, May 2024a. URL http://arxiv.org/abs/2301.10226. arXiv:2301.10226 [cs].

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. On the Reliability of Watermarks for Large Language Models, May 2024b. URL http://arxiv.org/abs/2306.04634. arXiv:2306.04634 [cs].

Arinbjorn Kolbeinsson, Kyle O'Brien, Tianjin Huang, Shanghua Gao, Shiwei Liu, Jonathan Richard Schwarz, Anurag Vaidya, Faisal Mahmood, Marinka Zitnik, Tianlong Chen, et al. Composable interventions for language models. *arXiv preprint arXiv:2407.06483*, 2024.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, et al. HunyuanVideo: A Systematic Framework For Large Video Generative Models, March 2025. URL http://arxiv.org/abs/2412.03603. arXiv:2412.03603 [cs].

Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Bhalerao, Christopher L. Buckley, Jason Phang, Samuel R. Bowman, and Ethan Perez. Pretraining language models with human preferences, 2023. URL https://arxiv.org/abs/2302.08582.

Tomek Korbak, Joshua Clymer, Benjamin Hilton, Buck Shlegeris, and Geoffrey Irving. A sketch of an AI control safety case, January 2025. URL http://arxiv.org/abs/2501.17315.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan Van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10: 50–72, 2022.

Rohith Kuditipudi, Jing Huang, Sally Zhu, Diyi Yang, Christopher Potts, and Percy Liang. Blackbox model provenance via palimpsestic membership inference. *arXiv preprint arXiv:2510.19796*, 2025.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, et al. The bigscience roots corpus: A 1.6 tb composite multilingual dataset." arxiv, 2023.

Bruce W. Lee, Addie Foote, Alex Infanger, Leni Shor, Harish Kamath, Jacob Goldman-Wetzler, Bryce Woodworth, Alex Cloud, and Alexander Matt Turner. Distillation Robustifies Unlearning, June 2025. URL http://arxiv.org/abs/2506.06278. arXiv:2506.06278 [cs].

Joon-Woo Lee, HyungChul Kang, Yongwoo Lee, Woosuk Choi, Jieun Eom, Maxim Deryabin, Eunsang Lee, Junghyun Lee, Donghoon Yoo, Young-Sik Kim, et al. Privacy-preserving machine learning with fully homomorphic encryption for deep neural network. *iEEE Access*, 10:30039–30054, 2022.

LG, Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Yemuk Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Kijeong Jeon, et al. Exaone deep: Reasoning enhanced language models. *arXiv preprint arXiv:2503.12524*, 2025.

Boheng Li, Renjie Gu, Junjie Wang, Leyi Qi, Yiming Li, Run Wang, Zhan Qin, and Tianwei Zhang. Towards resilient safety-driven unlearning for diffusion models against downstream fine-tuning. *arXiv preprint arXiv:2507.16302*, 2025a.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282, 2024a.

Kenneth Li, Yida Chen, Fernanda Viégas, and Martin Wattenberg. When bad data leads to good models. *arXiv preprint arXiv:2505.04741*, 2025b.

Linyang Li, Botian Jiang, Pengyu Wang, Ke Ren, Hang Yan, and Xipeng Qiu. Watermarking LLMs with Weight Quantization, October 2023. URL http://arxiv.org/abs/2310.11237. arXiv:2310.11237 [cs].

Shenghui Li, Edith C.-H. Ngai, Fanghua Ye, and Thiemo Voigt. PEFT-as-an-Attack! Jailbreaking Language Models during Federated Parameter-Efficient Fine-Tuning, December 2024b. URL http://arxiv.org/abs/2411.19335. arXiv:2411.19335 [cs].

Percy Liang, Rishi Bommasani, Kathleen A. Creel, and Rob Reich. The time is now to develop community norms for the release of foundation models, 2022. URL https://crfm.stanford.edu/2022/05/17/community-norms.html.

Yuqing Liang, Jiancheng Xiao, Wensheng Gan, and Philip S Yu. Watermarking techniques for large language models: A survey. *arXiv preprint arXiv:2409.00089*, 2024.

Zheng Lin, Xuanjie Hu, Yuxin Zhang, Zhe Chen, Zihan Fang, Xianhao Chen, Ang Li, Praneeth Vepakomma, and Yue Gao. Splitlora: A split parameter-efficient fine-tuning framework for large language models, 2024. URL https://arxiv.org/abs/2407.00952.

Johan Linåker, Cailean Osborne, Jennifer Ding, and Ben Burtenshaw. A cartography of open collaboration in open source ai: Mapping practices, motivations, and governance in 14 open large language model projects. *arXiv preprint arXiv:2509.25397*, 2025.

David Lindner and Mennatallah El-Assady. Humans are not Boltzmann Distributions: Challenges and Opportunities for Modelling Human Feedback and Interaction in Reinforcement Learning, June 2022. URL http://arxiv.org/abs/2206.13316. arXiv:2206.13316 [cs].

Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip S. Yu. A Survey of Text Watermarking in the Era of Large Language Models, August 2024a. URL http://arxiv.org/abs/2312.07913. arXiv:2312.07913 [cs].

Guozhi Liu, Weiwei Lin, Qi Mu, Tiansheng Huang, Ruichao Mo, Yuren Tao, and Li Shen. Targeted vaccine: Safety alignment for large language models against harmful fine-tuning via layer-wise perturbation. *IEEE Transactions on Information Forensics and Security*, 2025a.

Guozhi Liu, Qi Mu, Tiansheng Huang, Xinhua Wang, Li Shen, Weiwei Lin, and Zhang Li. Pharmacist: Safety alignment data curation for large language models against harmful fine-tuning. *arXiv preprint arXiv:2510.10085*, 2025b.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. Rethinking Machine Unlearning for Large Language Models, December 2024b. URL http://arxiv.org/abs/2402.08787. arXiv:2402.08787 [cs].

Xiaoqun Liu, Jiacheng Liang, Muchao Ye, and Zhaohan Xi. Robustifying Safety-Aligned Large Language Models through Clean Data Curation, May 2024c. URL http://arxiv.org/abs/2405.19358. arXiv:2405.19358 [cs].

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, et al. Llm360: Towards fully transparent open-source llms. *arXiv preprint arXiv:2312.06550*, 2023.

Michelle Lo, Shay B Cohen, and Fazl Barez. Large language models relearn removed concepts. *arXiv preprint arXiv:2401.01814*, 2024.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, et al. The data provenance initiative: A large scale audit of dataset licensing & attribution in ai. *arXiv preprint arXiv:2310.16787*, 2023a.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity, November 2023b. URL http://arxiv.org/abs/2305.13169. arXiv:2305.13169 [cs].

Shayne Longpre, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Blili-Hamelin, Yangsibo Huang, Aviya Skowron, Zheng-Xin Yong, Suhas Kotha, et al. A safe harbor for ai evaluation and red teaming. *arXiv preprint arXiv:2403.04893*, 2024a.

Shayne Longpre, Nikhil Singh, Manuel Cherep, Kushagra Tiwary, Joanna Materzynska, William Brannon, Robert Mahari, Naana Obeng-Marnu, Manan Dey, Mohammed Hamdy, et al. Bridging the data provenance gap across text, speech, and video. In *The Thirteenth International Conference on Learning Representations*, 2024b.

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 Adelani, Percy Liang, Rishi Bommasani, Peter Henderson, Sasha Luccioni, Yacine Jernite, and Luca Soldaini. The Responsible Foundation Model Development Cheatsheet: A Review of Tools & Resources, February 2025a. URL http://arxiv.org/abs/2406.16746. arXiv:2406.16746 [cs].

Shayne Longpre, Kevin Klyman, Ruth E. Appel, Sayash Kapoor, Rishi Bommasani, Michelle Sahar, Sean McGregor, Avijit Ghosh, Borhane Blili-Hamelin, Nathan Butters, et al. In-House Evaluation Is Not Enough: Towards Robust Third-Party Flaw Disclosure for General-Purpose AI, March 2025b. URL http://arxiv.org/abs/2503.16861. arXiv:2503.16861 [cs].

Peizhuo Lv, Pan Li, Shengzhi Zhang, Kai Chen, Ruigang Liang, Hualong Ma, Yue Zhao, and Yingjiu Li. A Robustness-Assured White-Box Watermark in Neural Networks. *IEEE Transactions on Dependable and Secure Computing*, 20(6):5214–5229, November 2023. ISSN 1941-0018. doi: 10.1109/TDSC.2023.3242737. URL https://ieeexplore.ieee.org/document/10038500.

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight methods to evaluate robust unlearning in llms. *arXiv preprint arXiv:2402.16835*, 2024.

Peihua Mai, Ran Yan, Zhe Huang, Youjia Yang, and Yan Pang. Split-and-denoise: Protect large language model inference with local differential privacy, 2024. URL https://arxiv.org/abs/2310.09130.

Emanuel Maiberg. Hugging face is hosting 5,000 nonconsensual ai models of real people, 7 2025. URL https://www.404media.co/hugging-face-is-hosting-5-000-nonconsensual-ai-models-of-real-people/.

Pratyush Maini, Sachin Goyal, Dylan Sam, Alex Robey, Yash Savani, Yiding Jiang, Andy Zou, Matt Fredrikson, Zacharcy C. Lipton, and J. Zico Kolter. Safety Pretraining: Toward the Next Generation of Safe AI, September 2025. URL http://arxiv.org/abs/2504.16980. arXiv:2504.16980 [cs].

Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, et al. Artificial intelligence index report 2025. *arXiv preprint arXiv:2504.07139*, 2025.

Sean McGregor, Allyson Ettinger, Nick Judd, Paul Albee, Liwei Jiang, Kavel Rao, William H Smith, Shayne Longpre, Avijit Ghosh, Christopher Fiorelli, et al. To err is ai: A case study informing llm flaw reporting practices. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28938–28945, 2025.

Sarah Meiklejohn, Hayden Blauzvern, Mihai Maruseac, Spencer Schrock, Laurent Simon, and Ilia Shumailov. Position: Machine learning models have a supply chain problem. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=zfohnbkMu0.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-Editing Memory in a Transformer, August 2023. URL http://arxiv.org/abs/2210.07229. arXiv:2210.07229 [cs].

Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal intelligence. Blog post, 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

Kyle Miller, Mia Hoffmann, and Rebecca Gelles. The use of open models in research. Research report, Center for Security and Emerging Technology, Georgetown University, October 2025. URL https://cset.georgetown.edu/publication/the-use-of-open-models-in-research/.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

Sella Nevo, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, and Jeff Alstott. Securing ai model weights. *Research reports, RAND*, 2024.

Helen Ngo, Cooper Raterink, João G. M. Araújo, Ivan Zhang, Carol Chen, Adrien Morisot, and Nicholas Frosst. Mitigating harm in language models with conditional-likelihood filtration, November 2021. URL http://arxiv.org/abs/2108.07790. arXiv:2108.07790 [cs].

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, March 2022. URL http://arxiv.org/abs/2112.10741. arXiv:2112.10741 [cs].

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

Ivica Nikolic, Teodora Baluta, and Prateek Saxena. Model Provenance Testing for Large Language Models, February 2025. URL http://arxiv.org/abs/2502.00706. arXiv:2502.00706 [cs].

Helen Nissenbaum. Accountability in a computerized society. *Science and engineering ethics*, 2(1):25–42, 1996.

Kyle O'Brien, Stephen Casper, Quentin Anthony, Tomek Korbak, Robert Kirk, Xander Davies, Ishan Mishra, Geoffrey Irving, Yarin Gal, and Stella Biderman. Deep Ignorance: Filtering Pretraining Data Builds Tamper-Resistant Safeguards into Open-Weight LLMs, August 2025. URL http://arxiv.org/abs/2508.06601. arXiv:2508.06601 [cs].

OpenAI. Chatgpt agent system card. Technical report, OpenAI, July 2025. URL https://cdn.openai.com/pdf/6bcccca6-3b64-43cb-a66e-4647073142d7/chatgpt_agent_system_card_launch.pdf.

OpenAI. Deep research system card. Technical report, OpenAI, February 2025. URL https://cdn.openai.com/deep-research-system-card.pdf. Accessed: 2025-10-07.

OpenAI. Disrupting malicious uses of ai: June 2025. Technical report, OpenAI Global Affairs, 2025a. URL https://openai.com/global-affairs/disrupting-malicious-uses-of-ai-june-2025/.

OpenAI. Sycophancy in gpt-4o: What happened and what we're doing about it. Web page, 2025b. URL https://openai.com/index/sycophancy-in-gpt-4o/.

Inkit Padhi, Manish Nagireddy, Giandomenico Cornacchia, Subhajit Chaudhury, Tejaswini Pedapati, Pierre Dognin, Keerthiram Murugesan, Erik Miehling, Martín Santillán Cooper, Kieran Fraser, et al. Granite guardian. *arXiv preprint arXiv:2412.07724*, 2024.

Giulio Pagnotta, Dorjan Hitaj, Briland Hitaj, Fernando Perez-Cruz, and Luigi V. Mancini. TATTOOED: A Robust Deep Neural Network Watermarking Scheme based on Spread-Spectrum Channel Coding, June 2024. URL http://arxiv.org/abs/2202.06091. arXiv:2202.06091 [cs].

Soumyadeep Pal, Changsheng Wang, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. Llm unlearning reveals a stronger-than-expected coreset effect in current benchmarks. *arXiv preprint arXiv:2504.10185*, 2025.

Jiadong Pan, Hongcheng Gao, Zongyu Wu, Taihang Hu, Li Su, Qingming Huang, and Liang Li. Leveraging catastrophic forgetting to develop safe diffusion models against malicious finetuning. *Advances in Neural Information Processing Systems*, 37:115208–115232, 2024.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2 (11), November 2021. ISSN 2666-3899. doi: 10.1016/j.patter.2021.100336. URL https://www.cell.com/patterns/abstract/S2666-3899(21)00184-7. Publisher: Elsevier.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

ShengYun Peng, Pin-Yu Chen, Matthew Hull, and Duen Horng Chau. Navigating the Safety Landscape: Measuring Risks in Finetuning Large Language Models, October 2024. URL http://arxiv.org/abs/2405.17374. arXiv:2405.17374 [cs].

Gabriel J Perin, Runjin Chen, Xuxi Chen, Nina ST Hirata, Zhangyang Wang, and Junyuan Hong. Lox: Low-rank extrapolation robustifies llm safety against fine-tuning. *arXiv preprint arXiv:2506.15606*, 2025.

Jason Phang, Herbie Bradley, Leo Gao, Louis Castricato, and Stella Biderman. Eleutherai: Going beyond" open science" to" science in the open". *arXiv preprint arXiv:2210.06413*, 2022.

Mary Phuong, Matthew Aitchison, Elliot Catt, Sarah Cogan, Alexandre Kaskasoli, Victoria Krakovna, David Lindner, Matthew Rahtz, Yannis Assael, Sarah Hodkinson, Heidi Howard, et al. Evaluating Frontier Models for Dangerous Capabilities, April 2024. URL http://arxiv.org/abs/2403.13793. arXiv:2403.13793 [cs].

Robert Podschwadt, Daniel Takabi, Peizhao Hu, Mohammad H Rafiei, and Zhipeng Cai. A survey of deep learning architectures for privacy-preserving machine learning with fully homomorphic encryption. *IEEE Access*, 10:117477–117500, 2022.

ATTACK VIA SAFETY PROJECTION. Spard: Defending harmful fine-tuning attack via safety projection with relevance–diversity data selection. 2025.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!, October 2023. URL http://arxiv.org/abs/2310.03693. arXiv:2310.03693 [cs].

Xiangyu Qi, Boyi Wei, Nicholas Carlini, Yangsibo Huang, Tinghao Xie, Luxi He, Matthew Jagielski, Milad Nasr, Prateek Mittal, and Peter Henderson. On Evaluating the Durability of Safeguards for Open-Weight LLMs, December 2024. URL http://arxiv.org/abs/2412.07097. arXiv:2412.07097 [cs].

C Raffel, N Shazeer, A Roberts, K Lee, S Narang, M Matena, Y Zhou, W Li, and PJ Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. arxiv preprint arxiv: 191010683. *Published online*, 2019.

Mohit Raghavendra, Vaskar Nath, and Sean Hendryx. Revisiting the Superficial Alignment Hypothesis, September 2024. URL http://arxiv.org/abs/2410.03717. arXiv:2410.03717 [cs].

Traian Rebedea, Razvan Dinu, Makesh Sreedhar, Christopher Parisien, and Jonathan Cohen. Nemo guardrails: A toolkit for controllable and safe llm applications with programmable rails. *arXiv preprint arXiv:2310.10501*, 2023.

Richard Ren, Steven Basart, Adam Khoja, Alexander Pan, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Gabriel Mukobi, Ryan H. Kim, Stephen Fitz, and Dan Hendrycks. Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress? *Advances in Neural Information Processing Systems*, 37:68559–68594, December 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7ebcdd0de471c027e67a11959c666d74-Abstract-Datasets_and_Benchmarks_Track.html.

Wei-Qing Ren, Yu-Ben Qu, Chao Dong, Yu-Qian Jing, Hao Sun, Qi-Hui Wu, and Song Guo. A survey on collaborative dnn inference for edge intelligence. *Machine Intelligence Research*, 20(3):370–395, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. arxiv 2022. *arXiv preprint arXiv:2112.10752*, 2021.

Domenic Rosati, Jan Wehner, Kai Williams, Lukasz Bartoszcze, David Atanasov, Robie Gonzales, Subhabrata Majumdar, Carsten Maple, Hassan Sajjad, and Frank Rudzicz. Representation Noising: A Defence Mechanism Against Harmful Finetuning, October 2024. URL http://arxiv.org/abs/2405.14577. arXiv:2405.14577 [cs].

Domenic Rosati, Sebastian Dionicio, Xijie Zeng, Subhabrata Majumdar, Frank Rudzicz, and Hassan Sajjad. Locking open weight models with spectral deformation. In *ICML Workshop on Technical AI Governance (TAIG)*, 2025.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 130136–130184. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/eb0b13cc515724ab8015bc978fdde0ad-Paper-Conference.pdf.

Debdeep Sanyal, Manodeep Ray, and Murari Mandal. Antidote: Bi-level adversarial training for tamper-resistant llms. *arXiv preprint arXiv:2509.08000*, 2025.

Rylan Schaeffer, Hailey Schoelkopf, Brando Miranda, Gabriel Mukobi, Varun Madan, Adam Ibrahim, Herbie Bradley, Stella Biderman, and Sanmi Koyejo. Why Has Predicting Downstream Capabilities of Frontier AI Models with Scale Remained Elusive?, February 2025. URL http://arxiv.org/abs/2406.04391. arXiv:2406.04391 [cs].

Paul S Scotti. How to structure open science collaborations online. 2025.

Elizabeth Seger. Open horizons: Exploring nuanced technical and policy approaches to openness in ai. Technical report, Demos in partnership with Mozilla, August 2024. URL https://demos.co.uk/wp-content/uploads/2024/08/Mozilla-Report_2024.pdf.

Elizabeth Seger and Jamie Hancock. The open dividend: Building an ai openness strategy to unlock the uk's ai potential. Technical report, Demos, June 2025. URL https://demos.co.uk/wp-content/uploads/2025/06/The-Open-Dividend_Report_2025.ac-2.pdf.

Elizabeth Seger, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K. Wei, Christoph Winter, Mackenzie Arnold, Sean O hEigeartaigh, Anton Korinek, et al. Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives, September 2023. URL http://arxiv.org/abs/2311.09227. arXiv:2311.09227 [cs].

Aqsa Shabbir, Halil İbrahim Kanpak, Alptekin Küpçü, and Sinem Sav. A taxonomy of attacks and defenses in split learning, 2025. URL https://arxiv.org/abs/2505.05872.

Mrinank Sharma, Meg Tong, Jesse Mu, Jerry Wei, Jorrit Kruthoff, Scott Goodfriend, Euan Ong, Alwin Peng, Raj Agarwal, Cem Anil, et al. Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming. *arXiv preprint arXiv:2501.18837*, 2025.

Han Shen, Pin-Yu Chen, Payel Das, and Tianyi Chen. Seal: Safety-enhanced aligned llm fine-tuning via bilevel data selection. *arXiv preprint arXiv:2410.07471*, 2024.

Hui Shen, Jingxuan Zhang, Boning Xiong, Rui Hu, Shoufa Chen, Zhongwei Wan, Xin Wang, Yu Zhang, Zixuan Gong, Guangyin Bao, et al. Efficient diffusion models: A survey. *arXiv preprint arXiv:2502.06805*, 2025.

Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Stephen Casper. Latent Adversarial Training Improves Robustness to Persistent Harmful Behaviors in LLMs, July 2025. URL http://arxiv.org/abs/2407.15549. arXiv:2407.15549 [cs].

Yunmeng Shu, Shaofeng Li, Tian Dong, Yan Meng, and Haojin Zhu. Model inversion in split learning for personalized llms: New insights from information bottleneck theory, 2025. URL https://arxiv.org/abs/2501.05965.

Vitaly Simonovich. Cato CTRL™ Threat Research: Uncovering Nytheon AI – A New Platform of Uncensored LLMs, June 2025. URL https://www.catonetworks.com/blog/cato-ctrl-nytheon-ai-a-new-platform-of-uncensored-llms/.

Stewart Slocum, Julian Minder, Clément Dumas, Henry Sleight, Ryan Greenblatt, Samuel Marks, and Rowan Wang. Believe it or not: How deeply do llms believe implanted facts? *arXiv preprint arXiv:2510.17941*, 2025.

Irene Solaiman. The Gradient of Generative AI Release: Methods and Considerations, February 2023. URL http://arxiv.org/abs/2302.04844. arXiv:2302.04844 [cs].

Irene Solaiman, Rishi Bommasani, Dan Hendrycks, Ariel Herbert-Voss, Yacine Jernite, Aviya Skowron, and Andrew Trask. Beyond Release: Access Considerations for Generative AI Systems, April 2025. URL http://arxiv.org/abs/250h2.16701. arXiv:2502.16701 [cs].

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*, 2024.

Filip Sondej and Yushi Yang. Collapse of irrelevant representations (cir) ensures robust and non-disruptive llm unlearning. *arXiv preprint arXiv:2509.11816*, 2025.

Madhulika Srikumar, Jiyoo Chang, Kasia Chmieliński, et al. Risk mitigation strategies for the open foundation model value chain. Technical Report Rev. 3.1, Partnership on AI, 2024. URL https://partnershiponai.org/wp-content/uploads/dlm_uploads/2024/07/open-foundation-model-risk-mitigation_rev3-1.pdf.

Marco Antonio Stranisci and Christian Hardmeier. What Are They Filtering Out? A Survey of Filtering Strategies for Harm Reduction in Pretraining Datasets, February 2025. URL http://arxiv.org/abs/2503.05721. arXiv:2503.05721 [cs].

Zekun Sun, Zhihao Sui, Na Ruan, Conghui He, Dahua Lin, and Jie LI. Trustworthy dataset proof: Certifying the authentic use of dataset in training models for enhanced trust. 2025.

Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, Andy Zou, Dawn Song, Bo Li, Dan Hendrycks, and Mantas Mazeika. Tamper-Resistant Safeguards for Open-Weight LLMs, February 2025. URL http://arxiv.org/abs/2408.00761. arXiv:2408.00761 [cs].

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025a.

GLM-4 5 Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. GLM-4.5: Agentic, Reasoning, and Coding (ARC) Foundation Models, August 2025b. URL http://arxiv.org/abs/2508.06471. arXiv:2508.06471 [cs].

National Telecommunications and Information Administration. Dual-use foundation models with widely available model weights. Technical report, NTIA / U.S. Department of Commerce, 2024. URL https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf.

Thorn and All Tech Is Human. Reducing the risk of synthetic content: Preventing generative AI from producing child sexual abuse material. Technical report, Thorn and All Tech Is Human, February 2024. URL https://www.nist.gov/system/files/documents/2024/02/15/ID012%20-%202024-02-01%2C%20Thorn%20and%20ATIH%2C%20Comments%20on%20AI%20EO%20RFI.pdf. Comments on AI Executive Order Request for Information submitted to NIST.

Johannes Treutlein, Dami Choi, Jan Betley, Sam Marks, Cem Anil, Roger Grosse, and Owain Evans. Connecting the Dots: LLMs can Infer and Verbalize Latent Structure from Disparate Training Data. *Advances in Neural Information Processing Systems*, 37:140667–140730, December 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/fe489a28a54583ee802b8e2955c024c2-Abstract-Conference.html.

UK AISI. Managing risks from increasingly capable open-weight ai systems. Blog post, 2025. URL https://www.aisi.gov.uk/blog/managing-risks-from-increasingly-capable-open-weight-ai-systems.

Emma Henderson Vaughan. NCMEC Releases New Data: 2024 in Numbers. URL https://ncmec.org/content/ncmec/en/blog/2025/ncmec-releases-new-data-2024-in-numbers.html.

Eric Wallace, Olivia Watkins, Miles Wang, Kai Chen, and Chris Koch. Estimating Worst-Case Frontier Risks of Open-Weight LLMs, August 2025. URL http://arxiv.org/abs/2508.03153. arXiv:2508.03153 [cs].

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, et al. Wan: Open and Advanced Large-Scale Video Generative Models, April 2025. URL http://arxiv.org/abs/2503.20314. arXiv:2503.20314 [cs].

Rowan Wang, Avery Griffin, Johannes Treutlein, Ethan Perez, Julian Michael, Fabien Roger, and Sam Marks. Modifying llm beliefs with synthetic document finetuning. Web post, 2025a. URL https://alignment.anthropic.com/2025/modifying-beliefs-via-sdf/.

Yuhui Wang, Rongyi Zhu, and Ting Wang. Self-destructive language model. *arXiv preprint arXiv:2505.12186*, 2025b.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. Assessing the Brittleness of Safety Alignment via Pruning and Low-Rank Modifications, October 2024. URL http://arxiv.org/abs/2402.05162. arXiv:2402.05162 [cs].

Boyi Wei, Zora Che, Nathaniel Li, Udari Madhushani Sehwag, Jasper GÃķtting, Samira Nedungadi, Julian Michael, Summer Yue, Dan Hendrycks, Peter Henderson, et al. Best practices for biorisk evaluations on open-weight bio-foundation models. *arXiv preprint arXiv:2510.27629*, 2025.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. Challenges in Detoxifying Language Models, September 2021. URL http://arxiv.org/abs/2109.07445. arXiv:2109.07445 [cs].

Eric Wu, Kevin Wu, and James Zou. Finetunebench: How well do commercial fine-tuning apis infuse knowledge into llms? *arXiv preprint arXiv:2411.05059*, 2024.

Suhang Wu, Ante Wang, Minlong Peng, Yujie Lin, Wenbo Li, Mingming Sun, and Jinsong Su. DocTER: Evaluating Document-based Knowledge Editing, July 2025. URL http://arxiv.org/abs/2308.09954. arXiv:2308.09954 [cs].

Yuxin Xiao, Sana Tonekaboni, Walter Gerych, Vinith Suriyakumar, and Marzyeh Ghassemi. When style breaks safety: Defending language models against superficial style alignment. *arXiv preprint arXiv:2506.07452*, 2025.

Zuan Xie, Yang Xu, Hongli Xu, Yunming Liao, and Zhiwei Yao. A novel hat-shaped device-cloud collaborative inference framework for large language models, 2025. URL https://arxiv.org/abs/2503.18989.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. Detoxifying Language Models Risks Marginalizing Minority Voices, April 2021. URL http://arxiv.org/abs/2104.06390. arXiv:2104.06390 [cs].

Xiaojun Xu, Yuanshun Yao, and Yang Liu. Learning to Watermark LLM-generated Text via Reinforcement Learning, March 2024. URL http://arxiv.org/abs/2403.10553. arXiv:2403.10553 [cs].

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, et al. Qwen3 Technical Report, May 2025. URL http://arxiv.org/abs/2505.09388. arXiv:2505.09388 [cs].

Mingke Yang, Yuqi Chen, Yi Liu, and Ling Shi. DistillSeq: A Framework for Safety Alignment Testing in Large Language Models using Knowledge Distillation, July 2024. URL http://arxiv.org/abs/2407.10106. arXiv:2407.10106 [cs] version: 1.

Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models, October 2023. URL http://arxiv.org/abs/2310.02949. arXiv:2310.02949 [cs].

Dixi Yao and Baochun Li. Is split learning privacy-preserving for fine-tuning large language models? *IEEE Transactions on Big Data*, pp. 1–12, 2024. doi: 10.1109/TBDATA.2024.3524101.

Biao Yi, Tiansheng Huang, Baolei Zhang, Tong Li, Lihai Nie, Zheli Liu, and Li Shen. Ctrap: Embedding collapse trap to safeguard large language models from harmful fine-tuning. *arXiv preprint arXiv:2505.16559*, 2025.

Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14428–14437, October 2021. doi: 10.1109/ICCV48922.2021.01418. URL https://ieeexplore.ieee.org/document/9711167. ISSN: 2380-7504.

Chen Yueh-Han, Nitish Joshi, Yulin Chen, Maksym Andriushchenko, Rico Angell, and He He. Monitoring Decomposition Attacks in LLMs with Lightweight Sequential Monitors, June 2025. URL http://arxiv.org/abs/2506.10949. arXiv:2506.10949 [cs].

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, Siyuan Cheng, Ziwen Xu, Xin Xu, Jia-Chen Gu, et al. A Comprehensive Study of Knowledge Editing for Large Language Models, November 2024. URL http://arxiv.org/abs/2401.01286. arXiv:2401.01286 [cs].

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, Baosong Yang, Chen Cheng, Jialong Tang, Jiandong Jiang, Jianwei Zhang, Jijie Xu, Ming Yan, Minmin Sun, Pei Zhang, Pengjun Xie, Qiaoyu Tang, Qin Zhu, Rong Zhang, Shibin Wu, Shuo Zhang, Tao He, Tianyi Tang, Tingyu Xia, Wei Liao, Weizhou Shen, Wenbiao Yin, Wenmeng Zhou, Wenyuan Yu, Xiaobin Wang, Xiaodong Deng, Xiaodong Xu, Xinyu Zhang, Yang Liu, Yeqiu Li, Yi Zhang, Yong Jiang, Yu Wan, and Yuxin Zhou. Qwen3guard technical report, 2025a. URL https://arxiv.org/abs/2510.14276.

Xuandong Zhao, Sam Gunn, Miranda Christ, Jaiden Fairoze, Andres Fabrega, Nicholas Carlini, Sanjam Garg, Sanghyun Hong, Milad Nasr, Florian Tramer, Somesh Jha, Lei Li, Yu-Xiang Wang, and Dawn Song. Sok: Watermarking for ai-generated content, 2025b. URL https://arxiv.org/abs/2411.18479.

Amber Yijia Zheng and Raymond A Yeh. Imma: Immunizing text-to-image models against malicious adaptation. In *European Conference on Computer Vision*, pp. 458–475. Springer, 2024.

Amber Yijia Zheng, Cedar Site Bai, Brian Bullins, and Raymond A Yeh. Model immunization from a condition number perspective. *arXiv preprint arXiv:2505.23760*, 2025.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions, September 2024. URL http://arxiv.org/abs/2305.14795. arXiv:2305.14795 [cs].

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less Is More for Alignment, May 2023. URL http://arxiv.org/abs/2305.11206. arXiv:2305.11206 [cs].

Kaiwen Zhou, Chengzhi Liu, Xuandong Zhao, Shreedhar Jangam, Jayanth Srinivasa, Gaowen Liu, Dawn Song, and Xin Eric Wang. The hidden risks of large reasoning models: A safety assessment of r1, 2025a. URL https://arxiv.org/abs/2502.12659.

Shiji Zhou, Lianzhe Wang, Jiangnan Ye, Yongliang Wu, and Heng Chang. On the limitations and prospects of machine unlearning for generative ai. *arXiv preprint arXiv:2408.00376*, 2024.

Yihe Zhou, Tao Ni, Wei-Bin Lee, and Qingchuan Zhao. A survey on backdoor threats in large language models (llms): Attacks, defenses, and evaluations. *arXiv preprint arXiv:2502.05224*, 2025b.

Sally Zhu, Ahmed Ahmed, Rohith Kuditipudi, and Percy Liang. Independence Tests for Language Models, March 2025a. URL http://arxiv.org/abs/2502.12292. arXiv:2502.12292 [cs].

Xiaochen Zhu, Xinjian Luo, Yuncheng Wu, Yangfan Jiang, Xiaokui Xiao, and Beng Chin Ooi. Passive inference attacks on split learning via adversarial regularization. In *Network and Distributed System Security (NDSS) Symposium 2025*, Feburary 2025b. doi: 10.14722/ndss.2025.23030.

Yuxuan Zhu, Tengjun Jin, Yada Pruksachatkun, Andy Zhang, Shu Liu, Sasha Cui, Sayash Kapoor, Shayne Longpre, Kevin Meng, Rebecca Weiss, et al. Establishing best practices for building rigorous agentic benchmarks. *arXiv preprint arXiv:2507.02825*, 2025c.

Daniel Ziegler, Seraphina Nix, Lawrence Chan, Tim Bauman, Peter Schmidt-Nielsen, Tao Lin, Adam Scherlis, Noa Nabeshima, Benjamin Weinstein-Raun, Daniel de Haas, Buck Shlegeris, and Nate Thomas. Adversarial training for high-stakes reliability. *Advances in Neural Information Processing Systems*, 35: 9274–9286, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/3c44405d619a6920384a45bce876b41e-Abstract-Conference.html.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving Alignment and Robustness with Circuit Breakers, July 2024. URL http://arxiv.org/abs/2406.04313. arXiv:2406.04313 [cs].

Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. An Adversarial Perspective on Machine Unlearning for AI Safety, May 2025. URL http://arxiv.org/abs/2409.18025. arXiv:2409.18025 [cs].