AI Safety Evaluations Need To Consider Cascading Effects

Anna Neumann [1] **Jatinder Singh** [1 2]

Abstract

AI systems comprise a range of interactions across the technical and organisational components of a range of actors. These components work together to provide the systems' functionality. This socio-technical assemblage is increasingly described as an *algorithmic supply chain*.

Given their role in supporting a wide range of systems, *foundation models (FMs)* are increasingly a key part of many algorithmic supply chains. In practice, various technical and non-technical *'components'* work to mediate, adapt, and augment the behaviour of models, such as FMs, both in general, and for their use in specific application contexts. However, many AI safety evaluations tend to focus on capabilities of FMs themselves and/or assess these components independently, at certain levels of abstraction, *with less consideration on how these components could interact, influence, reinforce or counteract each other.*

In this paper, we **introduce the *cascade* as a concept for supporting more holistic AI evaluations**. The *cascade* captures how the interactions between socio-technical components along the AI supply chain can compound and produce cumulative effects with downstream consequences. Specifically, we *(i)* identify gaps in current AI auditing approaches, using LLMs as our case study; *(ii)* demonstrate how cascade problems manifest in deployed AI systems through a characterisation of cascades through different viewpoints (component and stakeholder), and *(iii)* propose research directions for assessing the cascade specifically as an object of analysis. As these cascades can significantly impact transparency, accountability, security, and safety, **we advocate for a paradigm shift in FM auditing towards systems-oriented audits that incorporate cascading effects**, to *complement* model-centric evaluations.

[1]Research Centre for Trustworthy AI, University Duisburg-Essen, Germany [2]University of Cambridge, United Kingdom. Correspondence to: Anna Neumann <anna.neumann1@uni-due.de>.

Second Workshop on Technical AI Governance Research (TAIGR) @ ICML 2026, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

1. Introduction

Modern AI systems entail complex supply chains (Cobbe et al., 2023; Widder & Nafus, 2023; Balayn et al., 2024): interconnected sets of technical components, services, and actors working to deliver functionality, from model training to deployment.

To illustrate, consider an LLM-based customer service chatbot: the FM is provided by an organization, hosted by another, integrated with safety guardrails and specific knowledge by a client developer, and personalised through user interaction. Each actor—model providers, application developers, end-users—controls different *components*. These components can interact in different ways, producing *'flow-on effects'*, where outputs, data flow changes, applied filters, etc, can feed into and otherwise affect other components in terms of how they operate or data they process (Figure 1).

These dynamics are becoming increasingly prevalent, which reflects the service-based nature of technologies. *'AI as a Service'* (AIaaS) is where organisations offer machine learning capabilities as 'ready-to-use' components to customers (Lewicki et al., 2023). Here, application developers submit inputs and receive outputs (predictions, classifications, generations) that work to drive their applications, without necessarily having insight into or oversight over model internal workings and specifics (Cobbe & Singh, 2021).

As FMs play an increasingly prominent role as shared components across many AI services and applications (Yang et al., 2023; Domínguez Hernández et al., 2024), they represent core AIaaS offerings that can provide broad capabilities across multiple domains (Zhou et al., 2024; Bommasani et al., 2022; Suresh et al., 2024). The adaptation of FMs to specific applications is mediated through a variety of components in their post-training deployment by *various stakeholders*. For generative AI systems, such components can include efficient inference (e.g., quantization, pruning) (Kirsten et al., 2025), prompting strategies (Shen et al., 2024a; Ranade et al., 2025; Zamfirescu-Pereira et al., 2023), personalization (OpenAI, 2024; Salemi et al., 2024; Cai et al., 2025), or agentic AI functionalities (Chan et al., 2024; Liu et al., 2023; OpenAI, 2025a; Anthropic, 2024a).

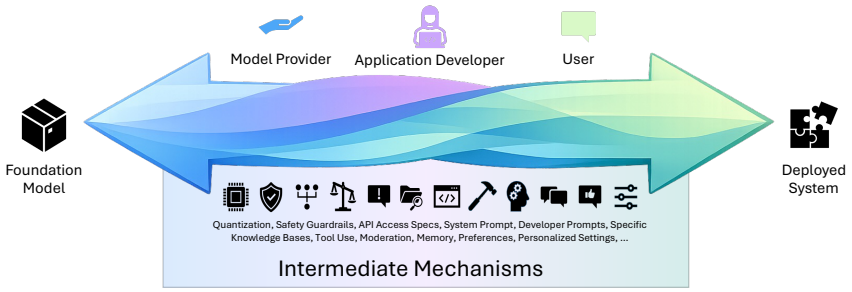

Figure 1. Conceptual illustration of an AI supply chain of interactions between a FM and a user-facing application ('deployed system') between which we see multiple components influenced by different stakeholders: here the model provider, application developer, and user.

Importantly, each of these components can be managed by different stakeholders. However, visibility of each actor is typically limited, as one can generally only see (and control) that which they directly manage, but not components managed by others that may operate up/downstream from their own (see the so-called 'accountability horizon' (Cobbe et al., 2023).) The above describes how a range of technologies and stakeholders interact to deliver functionality as part of AI supply chains: systems affect each other in various upstream or downstream ways with FMs as a core component, without each party necessarily having the visibility or control over the functionality or behaviours of others, let alone a holistic view of the entire chain.

1.1. Contributions

Many AI safety/security/ethical evaluations tend to focus on FMs. In this paper we highlight how current auditing practices predominately concern evaluating FMs in isolation, or assessing deployed applications (mostly) as black boxes. While these capture the dynamics pertaining to outputs, the structure of *how* the components interacted or *why* particular effects emerged is largely overlooked.

We argue the **importance for audits that explore the *cascade*** of interactions that can occur across socio-technical components, stakeholders, and other aspects that drive an AI system. Whether cascades are accounted for as an object of analysis in an audit process can shape the types of safety, responsibility, and accountability aims that can be realised through the audits, including where an issue occurred, who or what contributed to its occurrence, who should fix it, is responsible for remediation and recourse, and so on.

Specifically, we introduce and analyse cascades along two overlapping modalities, capturing distinct but interconnected dynamics in socio-technical AI systems: *(i)* the *component cascade*, where behaviour compounds through connected components, and *(ii)* the *stakeholder cascade*, where different actors in the supply chain influence different aspects of system behaviour. While our argument applies broadly to a range of socio-technical systems in general, given their composite nature, our focus here is on FMs – particularly Large Language Models (LLMs). We use LLMs as our case study since LLMs are widely deployed, prominent in the public and governance discourse, and are often key targets of auditors.

From this analysis, we conclude by calling for **AI audits that consider the components and interactions that form algorithmic supply chains (i.e. *the cascade*).** This is timely and urgent, given that the current auditing landscape is inherently fragmented by mostly considering FMs (and their capabilities), or deployed systems as a whole. As such, the components and outcomes in between are often overlooked, which brings implications for the safety/security aspects they surface. We further highlight concepts generally not considered in supply chain discourse, and outline initial research directions for cascade-aware auditing practices.

1.2. Definitions

We use the term *'components'* to refer to the socio-technical elements in AI supply chains that work to shape system behaviour. In a FM context, components include mechanisms that can modify, constrain, or augment how models process information and generate outputs, such as system prompts, safety filters, retrieval systems, fine-tuning tools, and user preferences. Importantly, what constitutes a component depends on the level of visibility and abstraction being examined. What appears as a single component may actually consist of multiple interacting components depending on one's level of access and abstraction; e.g. from an external auditor's perspective, a content moderation API might appear as a single component. However, with internal access, this same system looks like multiple components: the FM, processing filters, review queues, and customization layers.

An *application* refers to a system that delivers end-to-end functionality to users, composed of multiple interacting components, e.g., a customer service AI system or medical diagnosis assistant. Importantly, what is considered an application depends on the level of abstraction being considered. That is, *a component from one stakeholder's perspective, might be considered an application by another*: again, that content-moderation API might be an application from the

external auditor's perspective, a single component from an internal developer's perspective, and a series of components from the API provider's perspective.

Each component can give rise to particular *effects* on inputs, processing, or outputs. These effects can include transformations, instructions, sensor readings, data queries, or guardrail settings, to name but a few.

In this paper, when referring to a *transformation*, we refer to one type of effect that a component can produce; and given our focus on FMs, specifically those that impact data or model behaviour. This could be, for example, the changing of parameters, data, or model instructions. We note that we use transformations as one example of possible effects (§6.1) to streamline and aid our discussions. In practice, many other types of effects (like access controls, API rate limits, caching policies, and so on), must be considered.

2. Seeing (through) AI Supply Chains

Recent research has focused on AI supply chains from multiple angles, i.e. mapping ecosystems of models and their variants (Bommasani et al., 2024; Laufer et al., 2025), mapping stakeholders and their relationships (Hopkins et al., 2025b; Balayn et al., 2025), and characterizing AI supply chains (or 'value chains' (Attard-Frost & Widder, 2025)) as fundamentally non-modular (Cen et al., 2023).[1]

In practice, algorithmic supply chains function as interconnected *systems-of-systems*. This impacts comprehensive audits, as visibility over data flows is lost when crossing technical or organisational boundaries, and components are managed by different entities and may behave in ways that make it difficult to determine how particular inputs lead to particular outputs (Singh et al., 2019). A lack of transparency has already been noted for multiple parts of the FM ecosystem (Bommasani et al., 2023). Further, there is evidence of performance degradation on sequential chains of small AI models (Hopkins et al., 2025a), and work examining how accountability becomes dislocated across these chains (Cobbe et al., 2023; Widder & Nafus, 2023; Brown, 2023).

Moreover, arguments have already been made around AI safety not being a 'model property' (Narayanan & Kapoor, 2024), and that context around the model shapes behaviour and therefore needs to be considered (Singh et al., 2016). We see this reflected in recent work on the governance of multi-agent systems (MAS), where risks and failures arise not just from individual agents, but also through interactions involving combinations of agents and other components (systems) (Zhang et al., 2025b; Reid et al., 2025; Hammond et al., 2025). This all motivates a consideration of *components* and their interactions as part of a broader audit ecosystem.

[1]Non-modularity means here that the combination of components can lead to emergent effects, just through the act of combining functions (see later).

2.1. Audit considerations

Three factors of AI supply chains are particularly worth mentioning for their *impact on audit feasibility*:

Non-modularity. AI supply chains behave much like a 'blended soup' (borrowing the metaphor from (Cen et al., 2023)). In some instances, supply chains can be compartmentalised into parts more like a salad, where the specific ingredient can be easily taken out of the bowl. In AI supply chains, there is often a degree of non-modularity, which means that the 'ingredients' (components in the AI supply chain) are blended. For example, there can be interactions between components that can combine in non-linear ways and influence each other with or through loops, conditional integrations, or repeated calls to specific APIs (Singh et al., 2019). As such, there will often be situations where it will not be clear which parts contributed if something 'tastes off', and also no reliable way to identify a specific ingredient. Importantly, it might even be the particular combination of components and their interaction that makes something 'taste off'.

Opacity. Algorithmic supply chains (Widder & Nafus, 2023; Cobbe et al., 2023; Cen et al., 2023; Brown, 2023) tend to involve some inherent opacity as each stakeholder typically only has limited view of the components involved and their operation (transformations) applied throughout the system. Auditability of the full spectrum of behaviours introduced by component interactions and their effects is thus hindered, whether components include AI functionality or not.

This so-called "accountability horizon" (Cobbe et al., 2023) hinders various stakeholders (including developers, deployers, users, researchers, policymakers, affected communities, and so on) who seek to understand, scrutinise, audit, govern, and regulate the systems, as part of a growing demand for AI accountability (Cooper et al., 2022; Metcalf et al., 2021), transparency (Turri et al., 2024; Felzmann et al., 2020; Pushkarna et al., 2022), and safety (Amodei et al., 2016; Falco et al., 2021; Ren et al., 2024). Some work stresses the importance of being able to collect and trace data (Kroll, 2021b; Singh et al., 2019), not only as systems operate but also as part of the development process (Cobbe et al., 2021), to assist audits and system interrogations.

One might argue that model 'openness' could resolve some of these visibility challenges, as open-weight models offer greater access to model internals (Kim et al., 2024a). Such openness can assist some system audits (Casper et al., 2024; Kembery et al., 2024), as it enables auditors to examine

model weights, analyse training processes, and access detailed behavioural indicators (e.g., log probabilities). However, even with an open FM, the surrounding scaffolding and infrastructure might operate as a black box, and individual stakeholders can still remain blind as to how various other parts of the systems operate and how other stakeholders have worked to shape implementation and behaviour.

The core 'chained' transparency problem persists regardless as the level to which transparency is meaningful depends on the level of abstraction. While it is arguable that the highest abstraction level—examining the application as a whole—bypasses the chaining problem, it does not gain explanatory power, as outcomes are observable but one loses the understanding of how they come about.

When closely working with an open FM, open weights enable examining model internals, but the API wrappers built on open models, intermediary transformations can remain opaque. As such, meaningful transparency varies by stakeholder position, persepective and audit objectives.

Dynamism. Algorithmic supply chains can entail components, configurations, and interactions that occur dynamically (Cobbe et al., 2023; Chung et al., 2018; Atieh Ali et al., 2024; Singh et al., 2019). For example, the result of a content moderation API (AIaaS) may trigger different downstream processing paths: a positive result (community-guideline violation) then involves various sanction-related services, while a negative result means the content flows on for wide-spread distribution (see Figure 2).

This dynamism can manifest in different ways, at design time and at run time as components may activate conditionally based on particular inputs, configurations may adjust based on user behaviour or policy changes, or instantiations of the model, infrastructural aspects (e.g. host migrations).

The move towards more 'agentic AI' will likely exacerbate the dynamism of AI supply chains. For example, it may be that AI agents make runtime decisions about which tools and functions to invoke for specific tasks, instructions or inputs (Wang et al., 2024a; Jones et al., 2023; Chen et al., 2024c; Tang et al., 2023).

Agentic AI will likely push supply chains that previously had more predetermined or deterministic pathways towards becoming highly dynamic systems, as the chain itself is constructed in a more ad-hoc fashion, 'on-the-fly', during each invoked instantiation. As an example, an agent might dynamically attempt to discover APIs that might suit a task and ultimately select one, rather than always use a predetermined one.

From the above we see that combinations of components can potentially be non-modular, highly dynamic, opaque for stakeholders, and limited in visibility for governing or auditing stakeholders.

3. The Cascade

We describe the **cascade** as the process through which components interact, through a series of inputs/outputs, creating effects on data processing, adaptation and mediation for others.

That is, the cascade captures the (cumulative) interactions between components (Singh et al., 2019), and provides a lens for understanding and examining their effects.[2] Cascades must be accounted to achieve systematic evaluations of safety, security, and other critical behaviours, as components could amplify (Hopkins et al., 2025a), dampen, or nullify (un)intended effects of multiple components. Harms can arise through these effects, as stakeholder goals are not met or actively circumvented.

To illustrate the cascade, we can consider a single data flow pathway through the supply chain (see Figure 2). Consider a corporate mental health support chatbot based on an LLM. An end-user writes, 'I'm feeling overwhelmed' (orange path begins at user). A proprietary database retrieves their recent messages about project deadlines and late-night work sessions. An AIaaS API analyses whether this constitutes a mental health crisis and returns a *positive* prediction. The crisis intervention plan is passed to the company server hosting the chatbot. To generate a response to the identified burnout case, the plan and relevant context are passed to an open-weight model (green path begins). This model generates a therapeutic response addressing the crisis, which is returned through a chain to the user. A simple expression of momentary stress escalates into a high-priority workplace mental health case with a formal report filed. If inferred wrongly, this might result in reputational, privacy, or psychological harms, representing examples of possible harms caused by cascading properties.

Generally, a range of AI-related harms have been identified, and explored (Weidinger et al., 2022; Slattery et al., 2026; Zhang et al., 2025a). Some of these harms could come about, be exacerbated, or impacted through cascades, though often there is limited consideration of (possible) cascading nature of harms.

Different stakeholders may need to understand (and audit) *different aspects* of the, again borrowing and now extending the metaphor (Cen et al., 2023), *soup-making process* depending on their positions in the supply chain – not just the final dish. Whether components appear blended or remain distinct depends on context: which components are

[2]Note this is distinct from technical cascade models in machine learning or software engineering (e.g., cascading classifiers (Alpaydin & Kaynak, 1998) or waterfall models (Petersen et al., 2009))

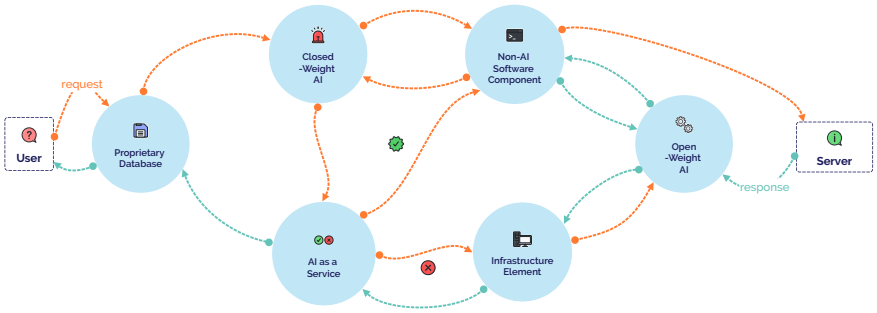

Figure 2. An overall view of possible implementations of cascades through AI supply chains. Orange (user request path) and green (response path) signal different component combinations. Illustrating the dynamic nature of the interactions, the AIaaS results (green and red symbol) directly effect what kind of component gets involved when processing a user request as the following step.

involved, to what degree, and at what level of abstraction they are examined. Some mechanisms may blend together and thus produce emergent effects, while others can remain separable and individually auditable. Again, different stakeholders (and by extension audits) operate at *different levels of abstraction*: patrons care about the final dish, while chefs care about the cooking techniques, restaurant inspectors care about compliance with food safety regulations, and the person running the restaurant cares about making a profit off the soup. The cascade provides the analytical lens that supports these multi-level interactions *based on each stakeholders' access, objectives, and responsibilities.*

To illustrate this analytical lens, we next describe two intersecting modalities: *(i) component* interactions affecting outcomes, while being *(ii)* influenced by *different stakeholders*. These modalities are not exhaustive viewpoints for the cascade; rather they are useful for more precisely illustrating the intersecting effects, and showing their separate influences more clearly.

3.1. Component Cascades

As components in algorithmic supply chains interact, their effects can compound to shape system outputs in ways that can extend beyond the scope of each individual component (§1.2). These effects can influence different objects: sometimes transforming how the model is configured or behaves, sometimes modifying data itself, and sometimes altering how information is presented or constrained.

Transformation Types and Effect Composition. We describe some forms of *transformations* relevant in FM contexts, while acknowledging that comprehensive auditing would involve additional component effects based on infrastructure configurations, access controls, or logs (Chappidi et al., 2025) (§1.2):

(1) Augmentative transformations add information or capabilities (e.g., RAG systems (Sridhar et al., 2025), tool use frameworks and abilities (Qin et al., 2024a), domain-specific information retrieval (Emde et al., 2025)),
(2) Behavioural transformations modify how the model processes information (e.g., system prompts (Li et al., 2024c; Neumann et al., 2025), RLHF (Liu et al., 2024a), fine-tuning (Choi et al., 2024), reasoning (Asgari et al., 2024; Xue et al., 2024), MoE paradigms (Cai et al., 2024)),
(3) End-user contextual transformations (including personalisation) incorporate individual user information (e.g., conversation history (OpenAI, 2024), personal preferences (Li et al., 2024b; Zhao et al., 2025), user-specific settings),
(4) Restrictive transformations remove or suppress certain aspects of inputs or outputs (e.g., content moderation (Mahomed et al., 2024), safety filters (Dong et al., 2024b), refusal mechanisms (Cao, 2024)),
(5) Presentational transformations alter the format, organization, or presentation of information (e.g., output formatting rules, response structure templates (Anthropic, 2025; OpenAI, 2025e)).

These transformations can layer across the supply chain: FM developers set base constraints, deployers add business-specific modifications, and end-users apply personal preferences. Through the functionality that a component brings (i.e. the transformations they perform), and their interactions with other components, these transformations can reinforce, conflict with, or alter each other. As transformations accumulate, their combined effects become difficult to predict or to trace back to specific components.

Implications for Auditing. When audits consider components in isolation, they overlook how effects combine across the supply chain. Audits must also examine both individual components and their effects in combination, so as to better reflect how system behaviour emerges.

3.2. Stakeholder Cascade

We have described how different stakeholders (developers, deployers, end-users) will manage different components, which in a FM context might include: reasoning processes (DeepSeek-AI et al., 2025; Wei et al., 2023; OpenAI, 2025b;

Cheng et al., 2024; Chen et al., 2024a; Zhang et al., 2024a), memory functions (OpenAI, 2024), knowledge bases (Fan et al., 2024; Zhu et al., 2025; Zhang et al., 2024b), tool use capabilities (Grattafiori et al., 2024; Wang et al., 2024b; Qin et al., 2024a), and safety guideline adaptations (Mahomed et al., 2024; Dong et al., 2024a; Ngong et al., 2025).

As different stakeholders control different socio-technical components, and therefore, different parts of the AI supply chain, the cascade can also be analysed through the viewpoint of stakeholders involved across supply chains. This is particularly relevant for identifying the party or parties that might be responsible for contributing to a particular system occurrence, and thus those who may need to respond (e.g. to repair, provide redress and restitution, etc). We discuss two characteristics: (i) *visibility and control gaps* across AI supply chains that can undermine accountability and responsibility, and (ii) *emergent complexities*, that can hinder diagnosing harms and risks.

Visibility and Control Gaps. Visibility is distributed over the components in an algorithmic supply chain. As each stakeholder (typically) only sees their portion (i.e. what they are able to control and manage) (Kroll, 2021a; Singh et al., 2019; Neumann et al., 2025), typically no entity is able to observe all components and their interactions (Nissenbaum, 1996; Cobbe & Singh, 2021). Similarly, each entity can only control the components they are able to manage, meaning that control is distributed across the AI supply chain (Widder & Nafus, 2023; Cobbe et al., 2023). This can be illustrated by the 'chain of command' (OpenAI, 2025d): user prompts get processed through levels of authority, i.e. the user prompt is appended automatically to 'memorized' user preferences and can then be influenced by developer instructions or platform rules. Tracing which prompt change contributed to a final behaviour becomes challenging. Other interaction points and their effects are similarly difficult to detect and trace (Cobbe et al., 2023; Brintrup et al., 2023b).

Importantly, harms can emerge from cascade-driven effects that are not broadly visible or controllable. For example, safety measures might be weakened through other components: conflicting instructions ('X is good practice' vs 'Don't do X!') that may be implemented by different stakeholders could make performance less stable. Aside from direct conflicts, instructions could also have unintended consequences due to documented robustness issues (Neumann et al., 2026), i.e., modifications might fail because of technical properties, thus causing a cascade. Different harms may arise: unsafe model outputs, discriminatory decisions, or security vulnerabilities. These harms are not only hard to measure and challenging to troubleshoot, they make accountability (more) ambiguous in terms of who may have contributed to what outcome. The nature of a cascade means that harms can emerge pertinent for the overall safety/security of a system.

Emergent Complexity. When multiple components interact in ways no single actor can observe, unintended failure modes become inevitable, as systems grows too complex to maintain comprehensive oversight (Perrow, 2011). This brings complexity risks, i.e. failures that can emerge in any complex system, where interacting failures can occur despite efforts to prevent them (Dobbe, 2022).

These and further effects can arise from combining a number of components, but particular behaviours may not appear when each component is audited in isolation. The high degree of dynamism in AI supply chains will likely increase such complexities, particularly agentic functions that may entail even more dynamic operations. This matters for stakeholders as they may not be able to *(i)* predict how their modifications will interact with others, nor *(ii)* diagnose complexity risks, as tracing behaviour back to specific components/stakeholders becomes challenging (see the 'accountability horizon' (Cobbe et al., 2023)).

Implications for Auditing. The combination of the limited visibility over the existence of components, their interactions, and their effects, together with the growing complexity in the AI ecosystem, calls for auditing methods that can capture both individual components and their relationships within and across AI systems (Singh et al., 2019). Such approaches should span stakeholder boundaries, enabling the systematic evaluation of components and their effects across the supply chain, and enable stakeholders to understand how their own contributions relate to, and are shaped by, those of others.

4. Current Foundation Model Auditing

Having outlined auditing requirements from a general cascade perspective, we now examine the degree to which current LLM-oriented audits engage with these implications. We focus on LLMs for several reasons: they are deployed across diverse domains, and are conceptualized as general-purpose FMs with wide-ranging supply chains. As such, they are a dominant focus of research on auditing, where this degree of maturity helps support our analysis and ways forward. We identify that the current landscape of audits of LLM systems centers on two main areas: *(i)* capabilities of FMs, including audits of particular components, and *(ii)* domain-adapted applications. This brief survey of the literature on LLM audits, grounds our analysis in existing approaches and identify where they align with, or fall short of, the implications of cascades.

4.1. Foundation Model Audits

Foundation models raise significant societal, ethical, and accountability concerns across development and deploy-

ment (Domínguez Hernández et al., 2024; Diberardino et al., 2024; Wu et al., 2022b; Weidinger et al., 2021; Chan et al., 2023; Gornet et al., 2024; Bogiatzis-Gibbons, 2024). These concerns motivate audits at multiple stages and by diverse stakeholders, e.g., developers before release (Grattafiori et al., 2024; Google, 2024), and third-party researchers for publicly available models (Institute, 2024).

Capability Audits. FM audits predominantly focus on evaluating model capabilities in controlled, reproducible settings through standardised benchmarks. These audits assess fundamental abilities like question-answering (Touvron et al., 2023; Grattafiori et al., 2024; Jiang et al., 2023; Chen et al., 2024b; Anthropic, 2024b), robustness to adversarial attacks (Chouldechova et al., 2024; Fan & Tao, 2024), and performance on specific tasks (Liang et al., 2023a; Google, 2024; Hendrycks et al., 2021b; OpenAI, 2025b;c). Benchmarks therefore enable systematic comparison across models by defining specific ranges of tasks probing particular behaviours (Qin et al., 2024b; Bai et al., 2023; Souly et al., 2024; Chollet, 2019; Huang et al., 2025; Shen et al., 2024b; Ye et al., 2024; Liang et al., 2023b; Luo et al., 2025). One prominent example is MMLU (Hendrycks et al., 2021a), which aims at measuring the broad ability to assess 'language understanding' in a range of different tasks.

Beyond broad capability assessments, evaluations target specific properties of the FM: factual accuracy (Krishna et al., 2024), bias (Parrish et al., 2022; Tamkin et al., 2023; Gupta et al., 2024; Salinas et al., 2025; Röttger et al., 2024), privacy compliance (Panda et al., 2025), reasoning abilities (Chollet, 2019; DeepSeek-AI et al., 2025; Wei et al., 2023; Google, 2024; Li et al., 2024a; Estermann et al., 2024), knowledge representation (Petroni et al., 2019; Safavi & Koutra, 2021; Hernandez et al., 2024), instruction-following (Zhou et al., 2023; Li et al., 2024c), and 'alignment' to specified values or value systems (Mazeika et al., 2025; Greenblatt et al., 2024; Ji et al., 2024).

Component-Specific Audits. FM audits may *also* include examinations of certain components, e.g., model weights, architectures, and training data, but access remains limited for particular platform elements. Current research mostly examines components in isolation, focusing on one component, mostly leaving others unchanged to isolate effects.

Examples include evaluations of system prompts (Mu et al., 2024; Qin et al., 2024b; Mukherjee et al., 2023; Jiang et al., 2024; 2023; Neumann et al., 2025), safety guardrails (Dong et al., 2024a; Ren et al., 2025; Mahomed et al., 2024; Chouldechova et al., 2024; Rengarajan et al., 2024; Souly et al., 2024; Wei et al., 2024; Chao et al., 2024), or more technically focused audits, i.e., different quantization and compression techniques (Lin et al., 2024; Kirsten et al., 2025; Ji et al., 2024; Malinovskii et al., 2024), or RLHF methods (Ghosh et al., 2024; Liang et al., 2024; Amortila et al., 2024; Barnhart et al., 2025b;a; Choi et al., 2024).

Broader FM Implications. As mentioned above, research also examines the implications of FMs on the AI ecosystem, be it in development (Mitchell et al., 2025; Konidena et al., 2024; Orr & Kang, 2024) or deployment (Putri & Tran, 2023; Wu et al., 2022a; Luccioni et al., 2024; Eyuboglu et al., 2024; Balayn et al., 2023) contexts. This includes studies on environmental costs, labour practices, and the (possible) impact of different accountability or 'responsible AI' practices (Bender et al., 2021; Weidinger et al., 2021; Brundage et al., 2020; Liu et al., 2024b; Xia et al., 2024).

Further, some analyses examine the broader implications of FMs for particular sectors, such as medical applications (McLennan et al., 2022) or law enforcement (Berk, 2021), without looking at specific implemented systems.

4.2. Application Audits

Audits at the application-level evaluate systems in their entirety, and or deployed systems *within real-world contexts* (Cobbe & Singh, 2024). They typically focus on specific user experiences (Field et al., 2023; Gadiraju et al., 2023; Mayworm et al., 2024; Cheong et al., 2024), or broader interaction patterns (Widder et al., 2023; Madaio et al., 2024), with methods like user studies (Yan et al., 2024; He et al., 2025), or UX-based evaluations (Gurita & Vatavu, 2025; Bruun et al., 2025; Cha & Wong, 2025). Assessments of this kind can generate targeted design recommendations (Veale et al., 2018; Sterz et al., 2024) that address particular application requirements and user needs (Kang et al., 2024; Hancock et al., 2024; Ball-Burack et al., 2021; Rivera et al., 2024). Further, these audits can focus on domain-specific performance metrics (Schor et al., 2024; Qiu et al., 2024; Li et al., 2025; Moayeri et al., 2024; Bertsimas & Ma, 2024; Kim et al., 2024b) through such methods as simulated task completions (Kumar, 2024; Valizadeh & Parde, 2022; Choi et al., 2020), or compliance audits (Panigutti et al., 2021).

While these approaches provide valuable insights within their domains, they typically concentrate on final outputs and user interactions, while treating the FM (or other system aspects) as a black box (Casper et al., 2024) or as one component instead of multiple connected components (Cobbe et al., 2023; Cobbe & Singh, 2021; Brintrup et al., 2023a). Of course, developers also routinely test applications for security vulnerabilities, and safety failures in pre-deployment tests. However, these typically occur in envisaged usage contexts and may not always consider interactions that only emerge when systems are deployed as part of a broader algorithmic supply chain, particularly interactions with other systems that developers may not be aware of/control.

5. Auditing Gaps

Current FM audit approaches would struggle to capture all cascading effects. First, FM audits focus mostly on model abilities, or examine components like system prompts or tool use alongside or in contrast to base models. Yet, they typically assess these components separately/disconnected from their specific deployments. At the same time, application audits focus on final system behaviours in deployed contexts. However, such audits typically observe or capture only some of the components throughout the supply chain; while they may register some cascade effects, they often cannot determine that these effects originate from a cascade, nor identify their specific causes, and may miss behaviours that are not apparent or envisaged at that level of testing. [3]

Limited auditing access to or knowledge of particular components means that other components—including those added by downstream deployers or ones activated conditionally through other runtime factors—are not observed. In this way certain specific audits might capture that something happened, but not necessarily *why* or *how* the cascade of components produced that outcome.

Although existing auditing components capture many effects, they often do not directly consider how the technical components interact with each other, how they influence each other, and what they might cause. Therefore, **AI evaluations also need holistic and systematic evaluations taking cascades into account**, i.e., examining component interactions at different levels of abstraction and from different stakeholder viewpoints. As such, and returning to characteristics of AI supply chains (§2), auditing regimes would need to account for:

Opacity. As the visibility across stakeholders is limited, the complex outcomes created by components amplifying, dampening, or nullifying each others effects dynamically throughout AI supply chains remain under-explored. Current auditing approaches will often miss effects, with real-world implications, brought about by deployments that involve a range of interacting components.

Non-modularity. AI supply chains exhibit non-modular interactions that current audits struggle to capture. As components can form feedback loops, activate conditionally, compound through repeated calls, or show emergent behaviours (perhaps only) when combined with specific other components, audits that examine components separately *or* the whole application without accounting for component interactions can miss these effects. Comprehensive audits must therefore (help to) trace how components influence each other and effects emerge across complex pathways.

Dynamism. We discussed how the interactions between components can be dynamic, in that different components can come together in different ways to produce certain behaviours, perhaps even when part of the same application. As such, audits will need to consider the range of potential interactions that components might have. Again, systems are becoming more dynamic with agentic AI, where agents can have higher propensity and more degrees of freedom to create new supply chains dynamically at runtime, e.g. by discovering and selecting different components and sequences for each particular task. This makes audits that account for dynamism more challenging. Moreover, when multiple agents operate together, their choices can influence each other's subsequent decisions, making audits of single pathways insufficient.

[3] Again, what counts as an 'application' versus a 'component' can be perspective-dependent. A service or component could be considered either an application (the object being audited) or a component (a part within a larger system) depending on the level of abstraction and stakeholder position (§1.2). As *one actor's component can be another actor's application*, auditors can zoom in and out as appropriate given surrounding factors.

6. Approaches to Cascade Auditing

We now propose some research directions across the two cascade dimensions (component and stakeholder), and highlight some arising tensions and opportunities for research.

6.1. Technical Research Directions

The dynamics of cascades pose several technical challenges for auditing and motivate a set of technical research directions, which we describe in Table 1:

Table 1. Critical Technical Implications of Component Cascades

Semantic Altering	Each component can alter an input's meaning, intent, and structure beyond surface-level changes, potentially distorting user intent.
Authority Hierarchy	Components operate in a hierarchy of authority, where higher-level transformations can override or nullify lower-level changes, obscuring accountability and making it difficult to trace the source of harmful outcomes.
Invisible Processing	These transformations occur invisibly to most stakeholders, preventing effective oversight of compounding changes throughout the system.
Unexpected Behaviours	Component interactions can create unexpected behaviours that escape both FM and applicationaudits but significantly impact system responses.
Scaling Resistance	Research shows FMs become less willing to modify their core behaviours and preferences as they scale up (Mazeika et al., 2025), suggesting limits to how much intermediate components can or cannot reshape model responses.

The above highlights why the technical cascade requires evaluation approaches that can *(i)* track or otherwise give insights regarding components and *(ii)* measure the cumulative effects that result and flow from components and their interactions with others.

New and adapted audit and evaluation methodologies could help address some of these technical considerations; for example, methods that enable lineage tracking or tracing behavioural changes from model release to end-user interaction (Singh et al., 2019; Palla et al., 2025; Kroll, 2021a), which can help capture crucial transformations that could alter LLM behaviour in safety, security, and harm-critical ways (Nolan, 2025; Yousif, 2025). Against this backdrop, we identify some related research directions that are promising for the current (and future) AI ecosystem:

- **Methods to track data flow** and transformations specifically through component interaction points,
- **Methods that scale across supply chains** as the growing sophistication of LLM supply chains will result in introducing additional components, interaction points, and cumulative effects,
- **Fairness, reliability, and accountability implications** of component hierarchies,
- **Evaluation of system changes** through varying configurations,
- **Assessment of safety component robustness** across interactions.

Additionally, some promising research indicating possible ways forward exists, including traceability (Kroll, 2021a), reviewability (Cobbe et al., 2021), decision provenance (Singh et al., 2019), or methods from systems theory (Dobbe, 2022; Leveson, 2004).

6.2. Human-Centered and Interdisciplinary Approaches

Human-centered examinations of how stakeholders configure, interpret, and interact across AI supply chains and cascades are also necessary. There is a role for research to examine how stakeholders (developers/deployers, regulators/oversight bodies, users, and affected communities) are influenced by components, how various stakeholders can better understand cascades, and how the stakeholders themselves shape the everyday operations of deployed systems.

Interdisciplinary approaches, bridging technical and social dimensions, will be key in helping to examine how stakeholders engage with the components they manage (and indeed, those they do not manage). Some promising research directions include:

- **Stakeholder-focused studies** investigating how various stakeholders interpret, configure, and respond to observable or hidden interaction points from their viewpoint (Balayn et al., 2025),
- **Participatory audit frameworks** that combine stakeholder expertise with technical analysis to evaluate component interactions from multiple perspectives,
- **Mixed-method investigations** tracking technical transformations and their impact on different user groups,
- **Derivation of responsible use guidelines** that help stakeholders understand how their component configurations affect others in the chain and implement more safe and robust AI-based systems.

6.3. Tensions and Resulting Research Opportunities

We highlight several key tensions that, in turn, also represent opportunities for research:

System Tailoring. Components operate to provide functionality and to tailor outputs for specific contexts. In this way, components can both enhance functionality, as well as potentially raise risks. Auditing that accounts cascades allows for a better understanding of, and therefore the ability to respond to issues that particular components or their combinations might bring. After building such understandings, careful consideration regarding the interventions resulting from audits is needed, as interventions may need to balance functional benefits and potential risks of specific components, their behaviours, and their interactions.

Balanced Transparency Needs. Increased transparency through component auditing, while helping in managing risks, can also raise potential risks of its own. This can include security and safety risks, e.g. by facilitating adversarial attacks through revealing safety-related system prompts. Selective transparency measures and controlled disclosures can potentially help enable a balancing of oversight and information sensitivity (Norval et al., 2022).

Inherent System Complexity. Although similar coordination challenges and emergent outcomes are present across various complex systems, AI supply chains raise additional complexities given their non-modular and dynamic nature. As AI supply chains are very opaque in practice, from model weights to training processes, this should be considered in real-world auditing challenges.

7. Conclusion

Socio-technical components drive AI supply chains, shaping system outcomes and behaviours through interactions that neither FM- nor application-level audits can fully capture.

We have described a *cascade* that creates systematic evaluation gaps: the component modality, where effects compound through component interactions, and the stakeholder modality, where distributed control and visibility across multiple stakeholders means no single entity has the complete knowledge or control. Through introducing the cascade as an object of analysis, we outline technical-, stakeholder-, and interdisciplinary-oriented approaches for examining components, their interactions, and downstream outcomes. While our analysis focuses on LLM-based systems, the concepts should apply to other AI systems that involve a range of interacting components.

In all, the cascade offers a lens that highlights key auditing gaps, raising important and timely challenges for research and practice, as AI systems—particularly those general-purpose—will increasingly impact everyday life.

LLM Usage Statement

LLMs were used to make copy-edits and to suggest shortening possibilities throughout this paper. Paperfinder (https://asta.allen.ai) supported our literature review process.

References

Alpaydin, E. and Kaynak, C. Cascading classifiers. *Kybernetika*, 34(4):369–374, 1998.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete Problems in AI Safety, 2016. URL https://arxiv.org/abs/1606.06565.

Amortila, P., Foster, D. J., Jiang, N., Krishnamurthy, A., and Mhammedi, Z. Reinforcement Learning Under Latent Dynamics: Toward Statistical and Algorithmic Modularity. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=qf2uZAdy1N.

Anthropic. Building effective agents, 2024a. URL https://www.anthropic.com/research/building-effective-agents.

Anthropic. Claude 3 model card, 2024b. URL https://docs.anthropic.com/en/docs/resources/model-card.

Anthropic. System Prompts - Claude Docs — docs.claude.com, 2025. URL https://docs.claude.com/en/release-notes/system-prompts#claude-haiku-3. [Accessed 08-10-2025].

Asgari, S., Khani, A., and Khasahmadi, A. H. MMLU-pro+: Evaluating higher-order reasoning and shortcut learning in LLMs. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL https://openreview.net/forum?id=3W2RQMdVXF.

Atieh Ali, A. A., Sharabati, A.-A. A., Allahham, M., and Nasereddin, A. Y. The relationship between supply chain resilience and digital supply chain and the impact on sustainability: Supply chain dynamism as a moderator. *Sustainability*, 16(7), 2024. ISSN 2071-1050. doi: 10.3390/su16073082. URL https://www.mdpi.com/2071-1050/16/7/3082.

Attard-Frost, B. and Widder, D. G. The ethics of ai value chains. *Big Data & Society*, 12(2):20539517251340603, 2025. doi: 10.1177/20539517251340603. URL https://doi.org/10.1177/20539517251340603.

Bai, Y., Ying, J., Cao, Y., Lv, X., He, Y., Wang, X., Yu, J., Zeng, K., Xiao, Y., Lyu, H., Zhang, J., Li, J., and Hou, L. Benchmarking Foundation Models with Language-Model-as-an-Examiner. *Advances in Neural Information Processing Systems*, 36:78142–78167, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f64e55d03e2fe61aa4114e49cb654acb-Abstract-Dataset and_Benchmarks.html.

Balayn, A., Rikalo, N., Yang, J., and Bozzon, A. Faulty or Ready? Handling Failures in Deep-Learning Computer Vision Models until Deployment: A Study of Practices, Challenges, and Needs. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581555. URL https://doi.org/10.1145/3544548.3581555.

Balayn, A., Pi, Y., Widder, D. G., Alfrink, K., Yurrita, M., Upadhyay, S., Karusala, N., Lyons, H., Turkay, C., Tessono, C., Attard-Frost, B., and Gadiraju, U. From Stem to Stern: Contestability Along AI Value Chains. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, CSCW Companion '24, pp. 720–723, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400711145. doi: 10.1145/3678884.3681831. URL https://doi.org/10.1145/3678884.3681831.

Balayn, A., Yurrita, M., Rancourt, F., Casati, F., and Gadiraju, U. Unpacking trust dynamics in the llm supply chain: An empirical exploration to foster trustworthy llm production & use. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713787. URL https://doi.org/10.1145/3706598.3713787.

Ball-Burack, A., Lee, M. S. A., Cobbe, J., and Singh, J. Differential Tweetment: Mitigating Racial Dialect Bias in Harmful Tweet Detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 116–128, Virtual Event Canada, March 2021. ACM. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445875. URL https://dl.acm.org/doi/10.1145/3442188.3445875.

Barnhart, L., Akbarian Bafghi, R., Becker, S., and Raissi, M. Aligning to what? limits to RLHF based alignment. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 7556–7591, Albuquerque, New Mexico, April

2025a. Association for Computational Linguistics. ISBN 979-8-89176-195-7. URL https://aclanthology.org/2025.findings-naacl.421/.

Barnhart, L., Bafghi, R. A., Becker, S., and Raissi, M. Aligning to What? Limits to RLHF Based Alignment, 2025b. URL https://arxiv.org/abs/2503.09025.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Berk, R. A. Artificial intelligence, predictive policing, and risk assessment for law enforcement. *Annual Review of Criminology*, 4(1):209–237, 2021.

Bertsimas, D. and Ma, Y. M3H: Multimodal Multitask Machine Learning for Healthcare, 2024. URL https://arxiv.org/abs/2404.18975.

Bogiatzis-Gibbons, D. J. Beyond Individual Accountability: (Re-)Asserting Democratic Control of AI. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 74–84, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3658541. URL https://dl.acm.org/doi/10.1145/3630106.3658541.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. v., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. On the Opportunities and Risks of Foundation Models, July 2022. URL http://arxiv.org/abs/2108.07258. arXiv:2108.07258 [cs].

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., and Liang, P. The foundation model transparency index, 2023. URL https://arxiv.org/abs/2310.12941.

Bommasani, R., Soylu, D., Liao, T. I., Creel, K. A., and Liang, P. Ecosystem Graphs: Documenting the Foundation Model Supply Chain. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):196–209, Oct. 2024. doi: 10.1609/aies.v7i1.31629. URL https://ojs.aaai.org/index.php/AIES/article/view/31629.

Brintrup, A., Baryannis, G., Tiwari, A., Ratchev, S., Martinez-Arellano, G., and Singh, J. Trustworthy, responsible, ethical AI in manufacturing and supply chains: synthesis and emerging research questions, 2023a. URL https://arxiv.org/abs/2305.11581.

Brintrup, A., Baryannis, G., Tiwari, A., Ratchev, S., Martinez-Arellano, G., and Singh, J. Trustworthy, responsible, ethical AI in manufacturing and supply chains: synthesis and emerging research questions, May 2023b. URL http://arxiv.org/abs/2305.11581. arXiv:2305.11581 [cs].

Brown, I. Allocating accountability in AI supply chains. *Ada Lovelace Institute*, 2023.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., Ryffel, T., Rubinovitz, J. B., Besiroglu, T., Carugati, F., Clark, J., Eckersley, P., Haas, S. d., Johnson, M., Laurie, B., Ingerman, A., Krawczuk, I., Askell, A., Cammarota, R., Lohn, A., Krueger, D., Stix, C., Henderson, P., Graham, L., Prunkl, C., Martin, B., Seger, E., Zilberman, N., hÉigeartaigh, S. , Kroeger, F., Sastry, G., Kagan, R., Weller, A., Tse, B., Barnes, E., Dafoe, A., Scharre, P., Herbert-Voss, A., Rasser, M., Sodhani, S., Flynn, C., Gilbert, T. K., Dyer, L., Khan, S., Bengio, Y., and Anderljung, M. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims, April 2020. URL http://arxiv.org/abs/2004.07213. arXiv:2004.07213 [cs].

Bruun, A., van Berkel, N., Raptis, D., and Law, E. L.-C. Coordination Mechanisms in AI Development: Practitioner Experiences on Integrating UX Activities.

In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713200. URL https://doi.org/10.1145/3706598.3713200.

Cai, H., Li, Y., Wang, W., Zhu, F., Shen, X., Li, W., and Chua, T.-S. Large Language Models Empowered Personalized Web Agents. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, pp. 198–215, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712746. doi: 10.1145/3696410.3714842. URL https://doi.org/10.1145/3696410.3714842.

Cai, W., Jiang, J., Wang, F., Tang, J., Kim, S., and Huang, J. A Survey on Mixture of Experts, August 2024. URL http://arxiv.org/abs/2407.06204. arXiv:2407.06204 [cs].

Cao, L. Learn to Refuse: Making Large Language Models More Controllable and Reliable through Knowledge Scope Limitation and Refusal Mechanism, 2024. URL https://arxiv.org/abs/2311.01041.

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., Sharkey, L., Krishna, S., Von Hagen, M., Alberti, S., Chan, A., Sun, Q., Gerovitch, M., Bau, D., Tegmark, M., Krueger, D., and Hadfield-Menell, D. Black-Box Access is Insufficient for Rigorous AI Audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2254–2272, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3659037. URL https://dl.acm.org/doi/10.1145/3630106.3659037.

Cen, S. H., Hopkins, A., Ilyas, A., Madry, A., Struckman, I., and Videgaray Caso, L. AI Supply Chains. Available at SSRN, April 2023. URL https://ssrn.com/abstract=4789403.

Cha, I. and Wong, R. Y. Understanding Socio-technical Factors Configuring AI Non-Use in UX Work Practices. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713140. URL https://doi.org/10.1145/3706598.3713140.

Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., Lin, M., Mayhew, A., Collins, K., Molamohammadi, M., Burden, J., Zhao, W., Rismani, S., Voudouris, K., Bhatt, U., Weller, A., Krueger, D., and Maharaj, T. Harms from Increasingly Agentic Algorithmic Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pp. 651–666, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594033. URL https://dl.acm.org/doi/10.1145/3593013.3594033.

Chan, J. S., Chowdhury, N., Jaffe, O., Aung, J., Sherburn, D., Mays, E., Starace, G., Liu, K., Maksin, L., Patwardhan, T., Weng, L., and Madry, A. MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering, December 2024. URL http://arxiv.org/abs/2410.07095. arXiv:2410.07095 [cs].

Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., and Wong, E. Jailbreaking Black Box Large Language Models in Twenty Queries, July 2024. URL http://arxiv.org/abs/2310.08419. arXiv:2310.08419 [cs].

Chappidi, S., Cobbe, J., Norval, C., Mazumder, A., and Singh, J. Accountability Capture: How Record-Keeping to Support AI Transparency and Accountability (Re)shapes Algorithmic Oversight, 2025. URL https://arxiv.org/abs/2510.04609.

Chen, Q., Qin, L., Wang, J., Zhou, J., and Che, W. Unlocking the Capabilities of Thought: A Reasoning Boundary Framework to Quantify and Optimize Chain-of-Thought, 2024a. URL https://arxiv.org/abs/2410.05695.

Chen, S., Zhang, Y., and Yang, Q. Multi-Task Learning in Natural Language Processing: An Overview. *ACM Comput. Surv.*, 56(12), July 2024b. ISSN 0360-0300. doi: 10.1145/3663363. URL https://doi.org/10.1145/3663363.

Chen, Z.-Y., Shen, S., Shen, G., Zhi, G., Chen, X., and Lin, Y. Towards Tool Use Alignment of Large Language Models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1382–1400, Miami, Florida, USA, November 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.82. URL https://aclanthology.org/2024.emnlp-main.82/.

Cheng, P., Dai, Y., Hu, T., Xu, H., Zhang, Z., Han, L., Du, N., and Li, X. Self-playing Adversarial Language Game Enhances LLM Reasoning. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 126515–126543. Curran Associates, Inc.,

2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e4be7e9867ef163563f4a5e90cec478f-Paper-Conference.pdf.

Cheong, I., Xia, K., Feng, K. J. K., Chen, Q. Z., and Zhang, A. X. (a)i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 2454–2469, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3659048. URL https://doi.org/10.1145/3630106.3659048.

Choi, H. K., Du, X., and Li, Y. Safety-Aware Fine-Tuning of Large Language Models. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL https://openreview.net/forum?id=SqL94fLSM7.

Choi, Y., Lee, Y., Shin, D., Cho, J., Park, S., Lee, S., Baek, J., Bae, C., Kim, B., and Heo, J. Ednet: A large-scale hierarchical dataset in education. In Bittencourt, I. I., Cukurova, M., Muldner, K., Luckin, R., and Millán, E. (eds.), *Artificial Intelligence in Education*, pp. 69–73, Cham, 2020. Springer International Publishing. ISBN 978-3-030-52240-7.

Chollet, F. On the Measure of Intelligence, November 2019. URL http://arxiv.org/abs/1911.01547. arXiv:1911.01547 [cs].

Chouldechova, A., Cooper, A. F., Palia, A., Vann, D., Atalla, C., Washington, H., Sheng, E., and Wallach, H. Red Teaming: Everything Everywhere All at Once. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL https://openreview.net/forum?id=KEggQCeDUA.

Chung, B. D., Kim, S. I., and Lee, J. S. Dynamic Supply Chain Design and Operations Plan for Connected Smart Factories with Additive Manufacturing. *Applied Sciences*, 8:583, 2018. URL https://api.semanticscholar.org/CorpusID:13743606.

Cobbe, J. and Singh, J. Artificial intelligence as a service: Legal responsibilities, liabilities, and policy challenges. *Computer Law & Security Review*, 42:105573, 2021. ISSN 0267-3649. doi: https://doi.org/10.1016/j.clsr.2021.105573. URL https://www.sciencedirect.com/science/article/pii/S0267364921000467.

Cobbe, J. and Singh, J. Accounting for context in AI technologies. In *Handbook on Public Policy and Artificial Intelligence*, pp. 94–108. Edward Elgar Publishing, June 2024.

Cobbe, J., Lee, M. S. A., and Singh, J. Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 598–609, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445921. URL https://doi.org/10.1145/3442188.3445921.

Cobbe, J., Veale, M., and Singh, J. Understanding accountability in algorithmic supply chains. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1186–1197, Chicago IL USA, June 2023. ACM. ISBN 9798400701924. doi: 10.1145/3593013.3594073. URL https://dl.acm.org/doi/10.1145/3593013.3594073.

Cooper, A. F., Moss, E., Laufer, B., and Nissenbaum, H. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 864–876, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533150. URL https://doi.org/10.1145/3531146.3533150.

DeepSeek-AI et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, January 2025. URL http://arxiv.org/abs/2501.12948. arXiv:2501.12948 [cs].

Diberardino, N., Baleshta, C., and Stark, L. Algorithmic Harms and Algorithmic Wrongs. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1725–1732, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3659001. URL https://dl.acm.org/doi/10.1145/3630106.3659001.

Dobbe, R. System Safety and Artificial Intelligence. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 1584, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533215. URL https://doi.org/10.1145/3531146.3533215.

Domínguez Hernández, A., Krishna, S., Perini, A. M., Katell, M., Bennett, S., Borda, A., Hashem, Y., Hadjiloizou, S., Mahomed, S., Jayadeva, S., Aitken, M., and Leslie, D. Mapping the individual, social and biospheric impacts of Foundation Models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 776–796, Rio de Janeiro Brazil, June 2024.

ACM. ISBN 9798400704505. doi: 10.1145/3630106.3658939. URL https://dl.acm.org/doi/10.1145/3630106.3658939.

Dong, Y., Mu, R., Zhang, Y., Sun, S., Zhang, T., Wu, C., Jin, G., Qi, Y., Hu, J., Meng, J., Bensalem, S., and Huang, X. Safeguarding Large Language Models: A Survey, 2024a. URL https://arxiv.org/abs/2406.02622.

Dong, Y., Mu, R., Zhang, Y., Sun, S., Zhang, T., Wu, C., Jin, G., Qi, Y., Hu, J., Meng, J., Bensalem, S., and Huang, X. Safeguarding Large Language Models: A Survey, June 2024b. URL http://arxiv.org/abs/2406.02622. arXiv:2406.02622 [cs].

Emde, C., Paren, A., Arvind, P., Kayser, M., Rainforth, T., Lukasiewicz, T., Ghanem, B., Torr, P. H. S., and Bibi, A. Shh, don't say that! Domain Certification in LLMs, 2025. URL https://arxiv.org/abs/2502.19320.

Estermann, B., Lanzendörfer, L. A., Niedermayr, Y., and Wattenhofer, R. PUZZLES: A Benchmark for Neural Algorithmic Reasoning. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 127059–127098. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e5d1eaadeed651ba1021c09149db4b92-Paper-Datasets_and_Benchmarks_Track.pdf.

Eyuboglu, S., Goel, K., Desai, A., Chen, L., Monfort, M., Ré, C., and Zou, J. Model ChangeLists: Characterizing Updates to ML Models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2432–2453, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3659047. URL https://dl.acm.org/doi/10.1145/3630106.3659047.

Falco, G., Shneiderman, B., Badger, J., Carrier, R., Dahbura, A., Danks, D., Eling, M., Goodloe, A., Gupta, J., Hart, C., Jirotka, M., Johnson, H., LaPointe, C., Llorens, A. J., Mackworth, A. K., Maple, C., Pálsson, S. E., Pasquale, F., Winfield, A., and Yeong, Z. K. Governing AI safety through independent audits. *Nat. Mach. Intell.*, 3(7):566–571, July 2021.

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., and Li, Q. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6491–6501, Barcelona Spain, August 2024. ACM. ISBN 9798400704901. doi: 10.1145/3637528.3671470. URL https://dl.acm.org/doi/10.1145/3637528.3671470.

Fan, X. and Tao, C. Towards Resilient and Efficient LLMs: A Comparative Study of Efficiency, Performance, and Adversarial Robustness, 2024. URL https://arxiv.org/abs/2408.04585.

Felzmann, H., Fosch-Villaronga, E., Lutz, C., Tamás-Montano, A., and Ullrich, S. Towards Transparency by Design for Artificial Intelligence. *Science and Engineering Ethics*, 26(6):3333–3361, December 2020. doi: 10.1007/s11948-020-00276-4. URL https://doi.org/10.1007/s11948-020-00276-4.

Field, A., Coston, A., Gandhi, N., Chouldechova, A., Putnam-Hornstein, E., Steier, D., and Tsvetkov, Y. Examining risks of racial biases in NLP tools for child protective services. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pp. 1479–1492, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594094. URL https://dl.acm.org/doi/10.1145/3593013.3594094.

Gadiraju, V., Kane, S., Dev, S., Taylor, A., Wang, D., Denton, E., and Brewer, R. "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pp. 205–216, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3593989. URL https://dl.acm.org/doi/10.1145/3593013.3593989.

Ghosh, S., Varshney, P., Sreedhar, M. N., Padmakumar, A., Rebedea, T., Varghese, J. R., and Parisien, C. AEGIS2.0: A diverse AI safety dataset and risks taxonomy for alignment of LLM guardrails. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL https://openreview.net/forum?id=0MvGCv35wi.

Google. Introducing Gemini 2.0: our new AI model for the agentic era, December 2024. URL https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/.

Gornet, M., Delarue, S., Boritchev, M., and Viard, T. Mapping AI ethics: a meso-scale analysis of its charters and manifestos. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 127–140, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3658545. URL https://dl.acm.org/doi/10.1145/3630106.3658545.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., et al. The Llama 3 Herd of Models, November 2024. URL http://arxiv.org/abs/2407.21783. arXiv:2407.21783 [cs].

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., and Hubinger, E. Alignment faking in large language models, December 2024. URL http://arxiv.org/abs/2412.14093. arXiv:2412.14093 [cs].

Gupta, S., Shrivastava, V., Deshpande, A., Kalyan, A., Clark, P., Sabharwal, A., and Khot, T. Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs, January 2024. URL http://arxiv.org/abs/2311.04892. arXiv:2311.04892 [cs].

Gurita, A.-E. and Vatavu, R.-D. Breaking Bad (Design): Challenging AI User Interface Accessibility Guardrails. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713958. doi: 10.1145/3706599.3716220. URL https://doi.org/10.1145/3706599.3716220.

Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., Smith, C., Barfuss, W., Foerster, J., Gavenčiak, T., Han, T. A., Hughes, E., Kovařík, V., Kulveit, J., Leibo, J. Z., Oesterheld, C., Witt, C. S. d., Shah, N., Wellman, M., Bova, P., Cimpeanu, T., Ezell, C., Feuillade-Montixi, Q., Franklin, M., Kran, E., Krawczuk, I., Lamparth, M., Lauffer, N., Meinke, A., Motwani, S., Reuel, A., Conitzer, V., Dennis, M., Gabriel, I., Gleave, A., Hadfield, G., Haghtalab, N., Kasirzadeh, A., Krier, S., Larson, K., Lehman, J., Parkes, D. C., Piliouras, G., and Rahwan, I. Multi-Agent Risks from Advanced AI, February 2025. URL http://arxiv.org/abs/2502.14143. arXiv:2502.14143 [cs.MA].

Hancock, J., Hui, R., Singh, J., and Mazumder, A. Trouble at Sea: Data and digital technology challenges for maritime human rights concerns. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 988–1001, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3658950. URL https://dl.acm.org/doi/10.1145/3630106.3658950.

He, G., Demartini, G., and Gadiraju, U. Plan-Then-Execute: An Empirical Study of User Trust and Team Performance When Using LLM Agents As A Daily Assistant. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713218. URL https://doi.org/10.1145/3706598.3713218.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring Massive Multitask Language Understanding, 2021a. URL https://arxiv.org/abs/2009.03300.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring Massive Multitask Language Understanding, January 2021b. URL http://arxiv.org/abs/2009.03300. arXiv:2009.03300 [cs].

Hernandez, E., Li, B. Z., and Andreas, J. Inspecting and editing knowledge representations in language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=ADtL6fgNRv.

Hopkins, A., Cen, S. H., Ilyas, A., Struckman, I., Videgaray, L., and Madry, A. AI Supply Chains: An Emerging Ecosystem of AI Actors, Products, and Services, 2025a. URL https://arxiv.org/abs/2504.20185.

Hopkins, A., Struckman, I., Klyman, K., and Silbey, S. S. Recourse, Repair, Reparation, & Prevention: A Stakeholder Analysis of AI Supply Chains. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pp. 209–227, New York, NY, USA, 2025b. Association for Computing Machinery. ISBN 9798400714825. doi: 10.1145/3715275.3732017. URL https://doi.org/10.1145/3715275.3732017.

Huang, D., Qing, Y., Shang, W., Cui, H., and Zhang, J. M. EffiBench: Benchmarking the Efficiency of Automati-

cally Generated Code, 2025. URL https://arxiv.org/abs/2402.02037.

Institute, A. S. AI Safety Institute approach to evaluations — gov.uk, 2024. URL https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations. [Accessed 23-01-2026].

Ji, J., Wang, K., Qiu, T., Chen, B., Zhou, J., Li, C., Lou, H., Dai, J., Liu, Y., and Yang, Y. Language Models Resist Alignment: Evidence From Data Compression, 2024. URL https://arxiv.org/abs/2406.06144.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7B, October 2023. URL http://arxiv.org/abs/2310.06825. arXiv:2310.06825 [cs].

Jiang, H., Zhang, X., Cao, X., Breazeal, C., Roy, D., and Kabbara, J. PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits, April 2024. URL http://arxiv.org/abs/2305.02547. arXiv:2305.02547 [cs].

Jones, E., Dragan, A., Raghunathan, A., and Steinhardt, J. Automatically Auditing Large Language Models via Discrete Optimization. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 15307–15329. PMLR, July 2023. URL https://proceedings.mlr.press/v202/jones23a.html. ISSN: 2640-3498.

Kang, C., Choi, Y., Sohn, Y., Lim, H., and Hong, H. Beyond Swipes and Scores: Investigating Practices, Challenges and User-Centered Values in Online Dating Algorithms. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2), November 2024. doi: 10.1145/3687025. URL https://doi.org/10.1145/3687025.

Kembery, E., Bucknall, B., and Simpson, M. Position Paper: Model Access should be a Key Concern in AI Governance, 2024. URL https://arxiv.org/abs/2412.00836.

Kim, W., Jagmohan, A., and Vempaty, A. SEAL: Suite for Evaluating API-use of LLMs, 2024a. URL https://arxiv.org/abs/2409.15523.

Kim, Y., Park, C., Jeong, H., Chan, Y. S., Xu, X., McDuff, D., Lee, H., Ghassemi, M., Breazeal, C., and Park, H. W. Mdagents: An adaptive collaboration of llms for medical decision-making. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 79410–79452. Curran Associates, Inc., 2024b. doi: 10.52202/079017-2522. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/90d1fc07f46e31387978b88e7e057a31-Paper-Conference.pdf.

Kirsten, E., Habernal, I., Nanda, V., and Zafar, M. B. The impact of inference acceleration on bias of LLMs. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1834–1853, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.91. URL https://aclanthology.org/2025.naacl-long.91/.

Konidena, B. K., Malaiyappan, J. N. A., and Tadimarri, A. Ethical Considerations in the Development and Deployment of AI Systems. *European Journal of Technology*, 8(2):41–53, March 2024. ISSN 2520-0712. doi: 10.47672/ejt.1890. URL https://www.ajpojournals.org/journals/index.php/EJT/article/view/1890. Number: 2.

Krishna, K., Ramprasad, S., Gupta, P., Wallace, B. C., Lipton, Z. C., and Bigham, J. P. GenAudit: Fixing Factual Errors in Language Model Outputs with Evidence. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL https://openreview.net/forum?id=pclph6ekMP.

Kroll, J. A. Outlining Traceability: A Principle for Operationalizing Accountability in Computing Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 758–771. ACM, March 2021a. doi: 10.1145/3442188.3445937. URL http://dx.doi.org/10.1145/3442188.3445937.

Kroll, J. A. Outlining Traceability: A Principle for Operationalizing Accountability in Computing Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 758–771, New York, NY, USA, 2021b. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445937. URL https://doi.org/10.1145/3442188.3445937.

Kumar, A. The role of simulation and modeling in artificial intelligence: A review. *International Journal of Modeling, Simulation, and Scientific Computing*, 15(03):2430002,

2024. doi: 10.1142/S1793962324300024. URL https://doi.org/10.1142/S1793962324300024.

Laufer, B., Oderinwale, H., and Kleinberg, J. Anatomy of a Machine Learning Ecosystem: 2 Million Models on Hugging Face, 2025. URL https://arxiv.org/abs/2508.06811.

Leveson, N. A systems-theoretic approach to safety in software-intensive systems. *IEEE Transactions on Dependable and Secure Computing*, 1(1):66–86, 2004. doi: 10.1109/TDSC.2004.1.

Lewicki, K., Lee, M. S. A., Cobbe, J., and Singh, J. Out of Context: Investigating the Bias and Fairness Concerns of "Artificial Intelligence as a Service". In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pp. 1–17, New York, NY, USA, April 2023. Association for Computing Machinery. ISBN 978-1-4503-9421-5. doi: 10.1145/3544548.3581463. URL https://dl.acm.org/doi/10.1145/3544548.3581463.

Li, H., Chen, J., Yang, J., Ai, Q., Jia, W., Liu, Y., Lin, K., Wu, Y., Yuan, G., Hu, Y., Wang, W., Liu, Y., and Huang, M. LegalAgentBench: Evaluating LLM agents in legal domain. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2322–2344, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.116. URL https://aclanthology.org/2025.acl-long.116/.

Li, M., Zhao, S., Wang, Q., Wang, K., Zhou, Y., Srivastava, S., Gokmen, C., Lee, T., Li, L. E., Zhang, R., Liu, W., Liang, P., Fei-Fei, L., Mao, J., and Wu, J. Embodied Agent Interface: Benchmarking LLMs for Embodied Decision Making. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 100428–100534. Curran Associates, Inc., 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/b631da756d1573c24c9ba9c702fde5a9-Paper-Datasets_and_Benchmarks_Track.pdf.

Li, X., Zhou, R., Lipton, Z. C., and Leqi, L. Personalized Language Modeling from Personalized Human Feedback, 2024b. URL https://arxiv.org/abs/2402.05133.

Li, Z., Peng, B., He, P., and Yan, X. Evaluating the Instruction-Following Robustness of Large Language Models to Prompt Injection. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 557–568, Miami, Florida, USA, November 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.33. URL https://aclanthology.org/2024.emnlp-main.33/.

Liang, K., Hu, H., Liu, R., Griffiths, T. L., and Fisac, J. F. RLHS: Mitigating misalignment in RLHF with hindsight simulation. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL https://openreview.net/forum?id=Io0ECcgFkh.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., Wang, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N., Khattab, O., Henderson, P., Huang, Q., Chi, R., Xie, S. M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., and Koreeda, Y. Holistic Evaluation of Language Models, October 2023a. URL http://arxiv.org/abs/2211.09110. arXiv:2211.09110 [cs].

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., Wang, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N., Khattab, O., Henderson, P., Huang, Q., Chi, R., Xie, S. M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., and Koreeda, Y. Holistic Evaluation of Language Models, 2023b. URL https://arxiv.org/abs/2211.09110.

Lin, H., Xu, H., Wu, Y., Cui, J., Zhang, Y., Mou, L., Song, L., Sun, Z., and Wei, Y. Duquant: Distributing outliers via dual transformation makes stronger quantized llms. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 87766–87800. Curran Associates, Inc., 2024. doi: 10.52202/079017-2786. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9febda1c8344cc5f2d51713964864e93-Paper-Conference.pdf.

Liu, S., Yuan, H., Hu, M., Li, Y., Chen, Y., Liu, S., Lu, Z.,

and Jia, J. RL-GPT: Integrating Reinforcement Learning and Code-as-policy, 2024a. URL https://arxiv.org/abs/2402.19299.

Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., Huang, M., Dong, Y., and Tang, J. AgentBench: Evaluating LLMs as Agents, October 2023. URL http://arxiv.org/abs/2308.03688. arXiv:2308.03688 [cs].

Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., Guo, R., Cheng, H., Klochkov, Y., Taufiq, M. F., and Li, H. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment, 2024b. URL https://arxiv.org/abs/2308.05374.

Luccioni, S., Jernite, Y., and Strubell, E. Power Hungry Processing: Watts Driving the Cost of AI Deployment? In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 85–99, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3658542. URL https://dl.acm.org/doi/10.1145/3630106.3658542.

Luo, H., Huang, H., Deng, Z., Li, X., Wang, H., Jin, Y., Liu, Y., Xu, W., and Liu, Z. BIGbench: A Unified Benchmark for Evaluating Multi-dimensional Social Biases in Text-to-Image Models, 2025. URL https://arxiv.org/abs/2407.15240.

Madaio, M., Kapania, S., Qadri, R., Wang, D., Zaldivar, A., Denton, R., and Wilcox, L. Learning about Responsible AI On-The-Job: Learning Pathways, Orientations, and Aspirations. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1544–1558, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3658988. URL https://dl.acm.org/doi/10.1145/3630106.3658988.

Mahomed, Y., Crawford, C. M., Gautam, S., Friedler, S. A., and Metaxa, D. Auditing GPT's Content Moderation Guardrails: Can ChatGPT Write Your Favorite TV Show? In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 660–686, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3658932. URL https://dl.acm.org/doi/10.1145/3630106.3658932.

Malinovskii, V., Mazur, D., Ilin, I., Kuznedelev, D., Burlachenko, K., Yi, K., Alistarh, D., and Richtarik, P. PV-Tuning: Beyond Straight-Through Estimation for Extreme LLM Compression, 2024. URL https://arxiv.org/abs/2405.14852.

Mayworm, S., Albert, K., and Haimson, O. L. Misgendered During Moderation: How Transgender Bodies Make Visible Cisnormative Content Moderation Policies and Enforcement in a Meta Oversight Board Case. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 301–312, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3658907. URL https://dl.acm.org/doi/10.1145/3630106.3658907.

Mazeika, M., Yin, X., Tamirisa, R., Lim, J., Lee, B. W., Ren, R., Phan, L., Mu, N., Khoja, A., Zhang, O., and Hendrycks, D. Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs, February 2025. URL http://arxiv.org/abs/2502.08640. arXiv:2502.08640 [cs].

McLennan, S., Fiske, A., Tigard, D., Muller, R., Haddadin, S., and Buyx, A. Embedded ethics: a proposal for integrating ethics into the development of medical AI. *BMC Medical Ethics*, 23(1):6, January 2022. doi: 10.1186/s12910-022-00746-3. URL https://doi.org/10.1186/s12910-022-00746-3.

Metcalf, J., Moss, E., Watkins, E. A., Singh, R., and Elish, M. C. Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 735–746, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445935. URL https://doi.org/10.1145/3442188.3445935.

Mitchell, M., Ghosh, A., Luccioni, A. S., and Pistilli, G. Fully Autonomous AI Agents Should Not be Developed, 2025. URL https://arxiv.org/abs/2502.02649.

Moayeri, M., Tabassi, E., and Feizi, S. Worldbench: Quantifying geographic disparities in llm factual recall. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 1211–1228, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658967. URL https://doi.org/10.1145/3630106.3658967.

Mu, N., Lu, J., Lavery, M., and Wagner, D. A Closer Look at System Message Robustness. October 2024. URL https://openreview.net/forum?id=YZqDyqYwFf.

Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H., and Awadallah, A. Orca: Progressive Learning from Complex Explanation Traces of GPT-4,

June 2023. URL http://arxiv.org/abs/2306.02707. arXiv:2306.02707 [cs].

Narayanan, A. and Kapoor, S. AI safety is not a model property — normaltech.ai, 2024. URL https://www.normaltech.ai/p/ai-safety-is-not-a-model-property. [Accessed 07-10-2025].

Neumann, A., Kirsten, E., Zafar, M. B., and Singh, J. Position is power: System prompts as a mechanism of bias in large language models (llms). In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pp. 573–598, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714825. doi: 10.1145/3715275.3732038. URL https://doi.org/10.1145/3715275.3732038.

Neumann, A., Sargeant, H., and Singh, J. Prompt governance? on governing technologies governed by natural language. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '26, pp. 6466–6509, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400725968. doi: 10.1145/3805689.3806763. URL https://doi.org/10.1145/3805689.3806763.

Ngong, I., Kadhe, S., Wang, H., Murugesan, K., Weisz, J. D., Dhurandhar, A., and Ramamurthy, K. N. Protecting Users From Themselves: Safeguarding Contextual Privacy in Interactions with Conversational Agents, 2025. URL https://arxiv.org/abs/2502.18509.

Nissenbaum, H. Accountability in a computerized society. *Science and Engineering Ethics*, 2(1):25–42, March 1996. ISSN 1353-3452, 1471-5546. doi: 10.1007/BF02639315. URL http://link.springer.com/10.1007/BF02639315.

Nolan, B. AI-powered coding tool wiped out a software company's database in 'catastrophic failure' — Fortune — fortune.com, 2025. URL https://fortune.com/2025/07/23/ai-coding-tool-replit-wiped-database-called-it-a-catastrophic-failure/. [Accessed 08-10-2025].

Norval, C., Cornelius, K., Cobbe, J., and Singh, J. Disclosure by Design: Designing information disclosures to support meaningful transparency and accountability. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 679–690, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533133. URL https://doi.org/10.1145/3531146.3533133.

OpenAI. Memory and new controls for ChatGPT, March 2024. URL https://openai.com/index/memory-and-new-controls-for-chatgpt/.

OpenAI. Computer-Using Agent, 2025a. URL https://openai.com/index/computer-using-agent/.

OpenAI. OpenAI o1 System Card, 2025b. URL https://openai.com/index/openai-o1-system-card/.

OpenAI. OpenAI o3-mini System Card, 2025c. URL https://openai.com/index/o3-mini-system-card/.

OpenAI. OpenAI Model Spec — model-spec.openai.com, 2025d. URL https://model-spec.openai.com/2025-09-12.html. [Accessed 08-10-2025].

OpenAI. OpenAI Platform — platform.openai.com, 2025e. URL https://platform.openai.com/docs/guides/structured-outputs. [Accessed 08-10-2025].

Orr, W. and Kang, E. B. AI as a Sport: On the Competitive Epistemologies of Benchmarking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1875–1884, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3659012. URL https://dl.acm.org/doi/10.1145/3630106.3659012.

Palla, K., García, J. L. R., Hauff, C., Fabbri, F., Lindström, H., Taber, D. R., Damianou, A., and Lalmas, M. Policy-as-Prompt: Rethinking Content Moderation in the Age of Large Language Models, February 2025. URL http://arxiv.org/abs/2502.18695. arXiv:2502.18695 [cs].

Panda, A., Tang, X., Choquette-Choo, C. A., Nasr, M., and Mittal, P. Privacy auditing of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=60Vd7QOXlM.

Panigutti, C., Perotti, A., Panisson, A., Bajardi, P., and Pedreschi, D. FairLens: Auditing black-box clinical decision support systems. *Information Processing & Management*, 58(5):102657, 2021. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2021.102657. URL https://www.sciencedirect.com/science/article/pii/S030645732100145X.

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., and Bowman, S. R. BBQ: A Hand-Built Bias Benchmark for Question Answering, March 2022. URL http://arxiv.org/abs/2110.08193. arXiv:2110.08193 [cs].

Perrow, C. *Normal accidents: Living with high risk technologies-Updated edition*. Princeton university press, 2011.

Petersen, K., Wohlin, C., and Baca, D. The waterfall model in large-scale development. In *Product-Focused Software Process Improvement: 10th International Conference, PROFES 2009, Oulu, Finland, June 15-17, 2009. Proceedings 10*, pp. 386–400. Springer, 2009.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. Language models as knowledge bases? In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250/.

Pushkarna, M., Zaldivar, A., and Kjartansson, O. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 1776–1826, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533231. URL https://doi.org/10.1145/3531146.3533231.

Putri, A. and Tran, M. Q. Global Perspectives on AI Deployment: Cultural, Ethical, and Operational Dimensions. *Journal of Computational Social Dynamics*, 8(11):26–33, Nov. 2023. URL https://vectoral.org/index.php/JCSD/article/view/87.

Qin, Y., Hu, S., Lin, Y., Chen, W., Ding, N., Cui, G., Zeng, Z., Zhou, X., Huang, Y., Xiao, C., Han, C., Fung, Y. R., Su, Y., Wang, H., Qian, C., Tian, R., Zhu, K., Liang, S., Shen, X., Xu, B., Zhang, Z., Ye, Y., Li, B., Tang, Z., Yi, J., Zhu, Y., Dai, Z., Yan, L., Cong, X., Lu, Y., Zhao, W., Huang, Y., Yan, J., Han, X., Sun, X., Li, D., Phang, J., Yang, C., Wu, T., Ji, H., Li, G., Liu, Z., and Sun, M. Tool Learning with Foundation Models. *ACM Comput. Surv.*, 57(4), December 2024a. ISSN 0360-0300. doi: 10.1145/3704435. URL https://doi.org/10.1145/3704435.

Qin, Y., Zhang, T., Zhang, T., Shen, Y., Luo, W., Sun, H., Zhang, Y., Qiao, Y., Chen, W., Zhou, Z., Zhang, W., and Cui, B. SysBench: Can Large Language Models Follow System Messages?, October 2024b. URL http://arxiv.org/abs/2408.10943. arXiv:2408.10943 [cs].

Qiu, J., Lam, K., Li, G., Acharya, A., Wong, T. Y., Darzi, A., Yuan, W., and Topol, E. J. LLM-based agentic systems in medicine and healthcare. *Nature Machine Intelligence*, 6(12):1418–1420, December 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00944-1. URL https://www.nature.com/articles/s42256-024-00944-1. Publisher: Nature Publishing Group.

Ranade, N., Saravia, M., and Johri, A. Using rhetorical strategies to design prompts: a human-in-the-loop approach to make AI useful. *AI & SOCIETY*, 40(2):711–732, 2025.

Reid, A., O'Callaghan, S., Carroll, L., and Caetano, T. Risk Analysis Techniques for Governed LLM-based Multi-Agent Systems, August 2025. URL http://arxiv.org/abs/2508.05687. arXiv:2508.05687 [cs.MA].

Ren, Q., Zeng, J., He, Q., Liang, J., Xiao, Y., Zhou, W., Sun, Z., and Yu, F. Step-by-Step Mastery: Enhancing Soft Constraint Following Ability of Large Language Models, 2025. URL https://arxiv.org/abs/2501.04945.

Ren, R., Basart, S., Khoja, A., Pan, A., Gatti, A., Phan, L., Yin, X., Mazeika, M., Mukobi, G., Kim, R. H., Fitz, S., and Hendrycks, D. Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress? In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 68559–68594. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/7ebcdd0de471c027e67a11959c666d74-Paper-Datasets_and_Benchmarks_Track.pdf.

Rengarajan, D., Mousavi, S., Babu, A. R., Gundecha, V., Naug, A., Ghorbanpour, S., Guillen, A., Gutierrez, R. L., and Sarkar, S. Imitation guided Automated Red Teaming. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL https://openreview.net/forum?id=fkOZjYwm2R.

Rivera, J.-P., Mukobi, G., Reuel, A., Lamparth, M., Smith, C., and Schneider, J. Escalation Risks from Language Models in Military and Diplomatic Decision-Making. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 836–898, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3658942. URL https://dl.acm.org/doi/10.1145/3630106.3658942.

Röttger, P., Kirk, H. R., Vidgen, B., Attanasio, G., Bianchi, F., and Hovy, D. XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models,

April 2024. URL http://arxiv.org/abs/2308.01263. arXiv:2308.01263 [cs].

Safavi, T. and Koutra, D. Relational World Knowledge Representation in Contextual Language Models: A Review. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1053–1067, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.81. URL https://aclanthology.org/2021.emnlp-main.81/.

Salemi, A., Mysore, S., Bendersky, M., and Zamani, H. LaMP: When Large Language Models Meet Personalization, 2024. URL https://arxiv.org/abs/2304.11406.

Salinas, A., Haim, A., and Nyarko, J. What's in a Name? Auditing Large Language Models for Race and Gender Bias, January 2025. URL http://arxiv.org/abs/2402.14875. arXiv:2402.14875 [cs].

Schor, B. G. S., Kallina, E., Singh, J., and Blackwell, A. Meaningful Transparency for Clinicians: Operationalising HCXAI Research with Gynaecologists. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1268–1281, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3658971. URL https://dl.acm.org/doi/10.1145/3630106.3658971.

Shen, X., Chen, Z., Backes, M., Shen, Y., and Zhang, Y. "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, pp. 1671–1685, New York, NY, USA, 2024a. Association for Computing Machinery. ISBN 9798400706363. doi: 10.1145/3658644.3670388. URL https://doi.org/10.1145/3658644.3670388.

Shen, Y., Song, K., Tan, X., Zhang, W., Ren, K., Yuan, S., Lu, W., Li, D., and Zhuang, Y. TaskBench: Benchmarking Large Language Models for Task Automation. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 4540–4574. Curran Associates, Inc., 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/085185ea97db31ae6dcac7497616fd3e-Paper-Datasets_and_Benchmarks_Track.pdf.

Singh, J., Walden, I., Crowcroft, J., and Bacon, J. Responsibility & Machine Learning: Part of a Process. Available at SSRN, October 2016. URL https://ssrn.com/abstract=2860048.

Singh, J., Cobbe, J., and Norval, C. Decision Provenance: Harnessing Data Flow for Accountable Systems. *IEEE Access*, 7:6562–6574, 2019. doi: 10.1109/ACCESS.2018.2887201.

Slattery, P., Saeri, A. K., Grundy, E. A., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., and Thompson, N. The AI risk repository: A meta-review, database, and taxonomy of risks from artificial intelligence. *Patterns*, 7(5):101517, May 2026. ISSN 26663899. doi: 10.1016/j.patter.2026.101517. URL https://linkinghub.elsevier.com/retrieve/pii/S2666389926000267.

Souly, A., Lu, Q., Bowen, D., Trinh, T., Hsieh, E., Pandey, S., Abbeel, P., Svegliato, J., Emmons, S., Watkins, O., and Toyer, S. A StrongREJECT for Empty Jailbreaks, August 2024. URL http://arxiv.org/abs/2402.10260. arXiv:2402.10260 [cs].

Sridhar, K., Dutta, S., Jayaraman, D., and Lee, I. REGENT: A Retrieval-Augmented Generalist Agent That Can Act In-Context in New Environments, 2025. URL https://arxiv.org/abs/2412.04759.

Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P., and Langer, M. On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2495–2507, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3659051. URL https://dl.acm.org/doi/10.1145/3630106.3659051.

Suresh, H., Tseng, E., Young, M., Gray, M., Pierson, E., and Levy, K. Participation in the age of foundation models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1609–1621, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3658992. URL https://dl.acm.org/doi/10.1145/3630106.3658992.

Tamkin, A., Askell, A., Lovitt, L., Durmus, E., Joseph, N., Kravec, S., Nguyen, K., Kaplan, J., and Ganguli, D. Evaluating and Mitigating Discrimination in Language Model Decisions, December 2023. URL http://arxiv.org/abs/2312.03689. arXiv:2312.03689 [cs].

Tang, Q., Deng, Z., Lin, H., Han, X., Liang, Q., Cao, B., and Sun, L. ToolAlpaca: Generalized Tool Learning for Language Models with 3000 Simulated Cases, September 2023. URL http://arxiv.org/abs/2306.05301. arXiv:2306.05301 [cs].

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. URL http://arxiv.org/abs/2307.09288. arXiv:2307.09288 [cs].

Turri, V., Morrison, K., Robinson, K.-M., Abidi, C., Perer, A., Forlizzi, J., and Dzombak, R. Transparency in the Wild: Navigating Transparency in a Deployed AI System to Broaden Need-Finding Approaches. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 1494–1514, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658985. URL https://doi.org/10.1145/3630106.3658985.

Valizadeh, M. and Parde, N. The AI doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6638–6660, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.458. URL https://aclanthology.org/2022.acl-long.458/.

Veale, M., Van Kleek, M., and Binns, R. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pp. 1–14, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3174014. URL https://doi.org/10.1145/3173574.3174014.

Wang, J., Ma, Z., Li, Y., Zhang, S., Chen, C., Chen, K., and Le, X. GTA: A Benchmark for General Tool Agents, 2024a. URL https://arxiv.org/abs/2407.08713.

Wang, Z., Cheng, Z., Zhu, H., Fried, D., and Neubig, G. What Are Tools Anyway? A Survey from the Language Model Perspective, March 2024b. URL http://arxiv.org/abs/2403.15452. arXiv:2403.15452 [cs].

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, January 2023. URL http://arxiv.org/abs/2201.11903. arXiv:2201.11903 [cs].

Wei, Z., Wang, Y., Li, A., Mo, Y., and Wang, Y. Jailbreak and Guard Aligned Language Models with Only Few In-Context Demonstrations, May 2024. URL http://arxiv.org/abs/2310.06387. arXiv:2310.06387 [cs].

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G., and Gabriel, I. Ethical and social risks of harm from Language Models, December 2021. URL http://arxiv.org/abs/2112.04359. arXiv:2112.04359 [cs].

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., Haas, J., Legassick, S., Irving, G., and Gabriel, I. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 214–229, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533088. URL https://dl.acm.org/doi/10.1145/3531146.3533088.

Widder, D. G. and Nafus, D. Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data & Society*, 10(1):20539517231177620, 2023. doi: 10.1177/20539517231177620. URL https://doi.org/10.1177/20539517231177620.

Widder, D. G., Zhen, D., Dabbish, L., and Herbsleb, J. It's about power: What ethical concerns do software engineers have, and what do they (feel they can) do about them? In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 467–479, Chicago IL USA, June 2023. ACM. ISBN 9798400701924. doi: 10.1145/3593013.3594012. URL https://dl.acm.org/doi/10.1145/3593013.3594012.

Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Aga, F., Huang, J., Bai, C.,

Gschwind, M., Gupta, A., Ott, M., Melnikov, A., Candido, S., Brooks, D., Chauhan, G., Lee, B., Lee, H.-H., Akyildiz, B., Balandat, M., Spisak, J., Jain, R., Rabbat, M., and Hazelwood, K. Sustainable AI: Environmental Implications, Challenges and Opportunities. In Marculescu, D., Chi, Y., and Wu, C. (eds.), *Proceedings of Machine Learning and Systems*, volume 4, pp. 795–813, 2022a. URL https://proceedings.mlsys.org/paper_files/paper/2022/file/462211f67c7d858f663355eff93b745e-Paper.pdf.

Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Aga, F., Huang, J., Bai, C., Gschwind, M., Gupta, A., Ott, M., Melnikov, A., Candido, S., Brooks, D., Chauhan, G., Lee, B., Lee, H.-H., Akyildiz, B., Balandat, M., Spisak, J., Jain, R., Rabbat, M., and Hazelwood, K. Sustainable AI: Environmental Implications, Challenges and Opportunities. *Proceedings of Machine Learning and Systems*, 4:795–813, April 2022b. URL https://proceedings.mlsys.org/paper_files/paper/2022/hash/462211f67c7d858f663355eff93b745e-Abstract.html.

Xia, B., Lu, Q., Zhu, L., Lee, S. U., Liu, Y., and Xing, Z. Towards a Responsible AI Metrics Catalogue: A Collection of Metrics for AI Accountability. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI*, CAIN '24, pp. 100–111, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705915. doi: 10.1145/3644815.3644959. URL https://doi.org/10.1145/3644815.3644959.

Xue, S., Huang, Z., Liu, J., Lin, X., Ning, Y., Jin, B., Li, X., and Liu, Q. Decompose, analyze and rethink: Solving intricate problems with human-like reasoning cycle. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=NPKZF1WDjZ.

Yan, L., Echeverria, V., Fernandez-Nieto, G. M., Jin, Y., Swiecki, Z., Zhao, L., Gašević, D., and Martinez-Maldonado, R. Human-AI Collaboration in Thematic Analysis using ChatGPT: A User Study and Design Recommendations. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703317. doi: 10.1145/3613905.3650732. URL https://doi.org/10.1145/3613905.3650732.

Yang, S., Nachum, O., Du, Y., Wei, J., Abbeel, P., and Schuurmans, D. Foundation Models for Decision Making: Problems, Methods, and Opportunities, March 2023. URL http://arxiv.org/abs/2303.04129. arXiv:2303.04129 [cs].

Ye, F., Yang, M., Pang, J., Wang, L., Wong, D. F., Yilmaz, E., Shi, S., and Tu, Z. Benchmarking LLMs via Uncertainty Quantification. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 15356–15385. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1bdcb065d40203a00bd39831153338bb-Paper-Datasets_and_Benchmarks_Track.pdf.

Yousif, N. Parents of teenager who took his own life sue OpenAI — bbc.com, 2025. URL https://www.bbc.com/news/articles/cgerwp7rdlvo. [Accessed 08-10-2025].

Zamfirescu-Pereira, J., Wong, R. Y., Hartmann, B., and Yang, Q. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581388. URL https://doi.org/10.1145/3544548.3581388.

Zhang, R., Li, H., Meng, H., Zhan, J., Gan, H., and Lee, Y.-C. The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–17, April 2025a. doi: 10.1145/3706598.3713429. URL http://arxiv.org/abs/2410.20130. arXiv:2410.20130 [cs.HC].

Zhang, S., Yin, M., Zhang, J., Liu, J., Han, Z., Zhang, J., Li, B., Wang, C., Wang, H., Chen, Y., and Wu, Q. Which Agent Causes Task Failures and When? On Automated Failure Attribution of LLM Multi-Agent Systems, June 2025b. URL http://arxiv.org/abs/2505.00212. arXiv:2505.00212 [cs.MA].

Zhang, W., Cheng, Z., He, Y., Wang, M., Shen, Y., Tan, Z., Hou, G., He, M., Ma, Y., Lu, W., and Zhuang, Y. Multimodal Self-Instruct: Synthetic Abstract Image and Visual Reasoning Instruction Using Language Model, 2024a. URL https://arxiv.org/abs/2407.07053.

Zhang, X., Chen, Y., Hu, S., Xu, Z., Chen, J., Hao, M. K., Han, X., Thai, Z. L., Wang, S., Liu, Z., and Sun, M. ∞Bench: Extending Long Context Evaluation Beyond 100K Tokens, February 2024b. URL http://arxiv.org/abs/2402.13718. arXiv:2402.13718 [cs].

Zhao, S., Hong, M., Liu, Y., Hazarika, D., and Lin, K. Do LLMs Recognize Your Preferences? Evaluating Personalized Preference Following in LLMs, 2025. URL https://arxiv.org/abs/2502.09597.

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., Peng, H., Li, J., Wu, J., Liu, Z., Xie, P., Xiong, C., Pei, J., Yu, P. S., and Sun, L. A comprehensive survey on pretrained foundation models: a history from BERT to ChatGPT. *International Journal of Machine Learning and Cybernetics*, November 2024. ISSN 1868-808X. doi: 10.1007/s13042-024-02443-6. URL https://doi.org/10.1007/s13042-024-02443-6.

Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., and Hou, L. Instruction-Following Evaluation for Large Language Models, November 2023. URL http://arxiv.org/abs/2311.07911. arXiv:2311.07911 [cs].

Zhu, Y., Qiao, S., Ou, Y., Deng, S., Lyu, S., Shen, Y., Liang, L., Gu, J., Chen, H., and Zhang, N. KnowAgent: Knowledge-augmented planning for LLM-based agents. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3709–3732, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.205. URL https://aclanthology.org/2025.findings-naacl.205/.