AI Safety Evaluations Need To Consider Cascading Effects

Anna Neumann
Research Centre Trust, UA Ruhr
University of Duisburg-Essen, Germany
anna.neumann1@uni-due.de

Jatinder Singh
Research Centre Trust, UA Ruhr
University of Duisburg-Essen, Germany &
University of Cambridge, United Kingdom
jatinder.singh@uni-due.de

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 AI system auditing towards systems-oriented audits that incorporate cascading effects**, to *complement* model centric evaluations.

1 Introduction

Modern AI systems entail complex supply chains [1; 2; 3]: 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 personalized 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 the data they process (see Figure 1).

Article presented at the Second Conference of the International Association for Safe and Ethical Artificial Intelligence (IASEAI 2026).

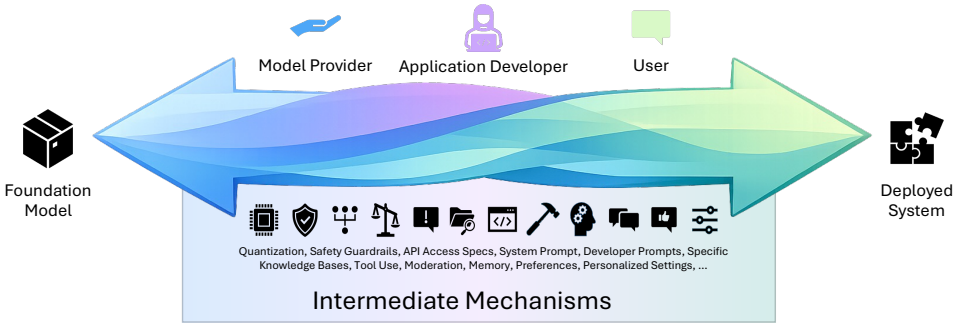

Figure 1: A conceptual illustration an AI supply chain depicting the interactions between a FM and a user-facing application ('deployed system') between which we see multiple components influenced by different stakeholders: here the model provider, application developer, and user.

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 [4]. 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 [5].

As FMs play an increasingly prominent role as shared components across many AI services and applications [6; 7], they represent core AIaaS offerings that can provide broad capabilities across multiple domains [8; 9; 10]. 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) [11], prompting strategies [12; 13; 14], personalization [15; 16; 17], or agentic AI functionalities [18; 19; 20; 21].

Importantly, each of these components can be potentially managed by different stakeholders. However, the visibility of each actor is typically limited, as one can generally only see (and control) that which they directly manage, but not those components managed by others that may operate up/downstream from their own.

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 technical/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 directly 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, who 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, post-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, routing, and so on), have to be considered.

2 Background: 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 [22; 23], mapping stakeholders and their relationships [24; 25], and characterizing AI supply chains (or 'value chains' [26]) as fundamentally non-modular [27]. Non-modularity means here that the combination of components can lead to emergent effects, just through the act of combining functions (see below).

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 [28]. A lack of transparency has already been noted for multiple parts of the FM ecosystem [29]. Further, there is evidence of performance degradation on sequential chains of small AI models [30], and work examining how accountability becomes dislocated across these chains [1; 2; 31]. Moreover, arguments have already been made around AI safety not being a 'model property' [32], and that context around the model shapes behaviour and therefore needs to be considered [33]. This all motivates our consideration of *components* and their interactions as part of a broader audit ecosystem. 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 [27]). 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 [28]. 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 [2; 1; 27; 31] tend to involve some inherent opacity as each stakeholder typically only has limited view of the components involved and the 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. The limited visibility across supply chains 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 [1; 34; 35], transparency [36; 37; 38], and safety [39; 40; 41]. Some work stresses the importance of being able to collect and trace data [42; 28], not only as systems operate but also as part of the development process [43], 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 [44]. Such openness can assist some system audits [45; 46], 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 [1; 47; 48]. 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), etc.

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 [49; 50; 51; 52]. 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 pre-determined 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 [28], and provides a lens for

understanding and examining their effects.[1] Cascades must be accounted to achieve systematic evaluations of safety, security, and other critical behaviours, as components could amplify [30], 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 analyzes 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.

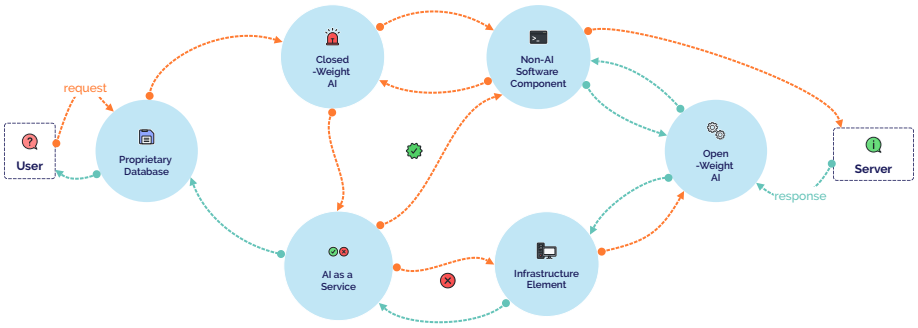

Figure 2: An overall view of possible implementations of cascades through AI supply chains. Orange (user request path) and green (response path) signal different component combinations. Illustrating the dynamic nature of the interactions, the AIaaS results (green and red symbol) directly effect what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

Different stakeholders may need to understand (and audit) *different aspects* of the, again borrowing and now extending the metaphor[27], *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 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.

[1]Note this is distinct from technical cascade models in machine learning or software engineering (e.g., cascading classifiers [53] or waterfall models [54])

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 [55] (§1.2):

(1) Augmentative transformations add information or capabilities (e.g., RAG systems [56], tool use frameworks and abilities [57], domain-specific information retrieval [58]),
(2) Behavioural transformations modify how the model processes information (e.g., system prompts [59; 60], RLHF [61], fine-tuning [62], reasoning [63; 64], Mixture-of-Experts paradigms [65]),
(3) End-user contextual transformations (including personalisation) incorporate individual user information (e.g., conversation history [15], personal preferences [66; 67], user-specific settings),
(4) Restrictive transformations remove or suppress certain aspects of inputs or outputs (e.g., content moderation [68], safety filters [69], refusal mechanisms [70]),
(5) Presentational transformations alter the format, organization, or presentation of information (e.g., output formatting rules, response structure templates [71; 72]).

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 [73; 74; 75; 76; 77; 78], memory functions [15], knowledge bases [79; 80; 81], tool use capabilities [82; 83; 57], and safety guideline adaptations [68; 84; 85].

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) [86; 28; 60], typically no entity is able to observe all components and their interactions [87; 5]. Similarly, each entity can only control the components they are able to manage, meaning that control is distributed across the AI supply chain [2; 1]. This can be illustrated by the 'chain of command' [88] component: user prompts get processed through levels of authority, i.e. the user prompt is appended automatically by stored 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 [5; 1; 89].

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 modifications ('X is good practice' vs 'Don't do X!') could make the model less stable, or modifications dependent on multiple parties might fail. This could result in different harms: 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. Ultimately, the nature of a cascade means that risks 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 [90]. This brings complexity risks, i.e. failures that can emerge in any complex system, where interacting failures can occur despite efforts to prevent them [91].

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' [1]).

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 [28]. 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 The Current Foundation Model AI Auditing Landscape

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 their development and deployment [7; 92; 93; 94; 95; 96; 97]. These concerns motivate audits at multiple stages and by diverse stakeholders, e.g., developers test models before release [82; 98], and third-party researchers evaluate publicly available models [99].

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 [100; 82; 101; 102; 103], robustness to adversarial attacks [104; 105], and performance on specific tasks [106; 98; 107; 75; 108]. Benchmarks therefore enable systematic comparison across models by defining specific ranges of tasks probing particular behaviours [109; 110; 111; 112; 113; 114; 115; 116; 117]. One prominent example is MMLU [118], 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 [119], bias [120; 121; 122; 123; 124], privacy compliance [125], reasoning abilities [112; 73; 74; 98; 126; 127], knowledge representation [128; 129; 130], instruction-following [131; 59], and 'alignment' to specified values or value systems [132; 133; 134].

Component-Specific Audits. FM audits may *also* include examinations of certain technical 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 [135; 109; 136; 137; 101; 60], safety guardrails [84; 138; 68; 104; 139; 111; 140; 141], or more technically focused audits, i.e., different quantization and compression techniques [142; 11; 134; 143], or RLHF methods [144; 145; 146; 147; 148; 62].

Broader FM Implications. As mentioned above, research also examines the implications of FMs on the AI ecosystem, be it in development [149; 150; 151] or deployment [152; 153; 154; 155; 156] contexts. This includes studies on environmental costs, labour practices, and the (possible) impact of

different accountability or 'responsible AI' practices [157; 94; 158; 159; 160].Further, some analyses examine the broader implications of FMs for particular sectors, such as medical applications [161] or law enforcement [162], 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* [163]. They typically focus on specific user experiences [164; 165; 166; 167], or broader interaction patterns [168; 169], with methods like user studies [170; 171], or UX-based evaluations [172; 173; 174]. Assessments of this kind can generate targeted design recommendations [175; 176] that address particular application requirements and user needs [177; 178; 179; 180]. Further, these audits can focus on domain-specific performance metrics [181; 182; 183; 184; 185; 186] through such methods as simulated task completions [187; 188; 189], or compliance audits [190].

While these approaches provide valuable insights within their domains, they typically concentrate on final outputs and user interactions, while treating the FM (or other aspects of the system) as a black box [45] or as one component instead of multiple connected components [1; 5; 191]. 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.[2] 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 how effects emerge across complex pathways.

[2]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.

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.

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 application-level audits but significantly impact system responses.
Scaling Resistance	Research shows that FMs become less willing to modify their core behaviours and preferences as they scale up [132], suggesting fundamental 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 [28; 192; 86], which can help capture crucial transformations that could alter LLM behaviour in safety, security, and harm-critical ways [193; 194].

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 interaction points.

Additionally, some promising research indicating possible ways forward exists, including traceability [86], reviewability [43], decision provenance [28], or methods from systems theory [91; 195].

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 various 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 specific viewpoint (see [25; 196]),
- **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 supply 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 their own. This can include security and safety risks, e.g. by facilitating adversarial attacks, such as where revealing safety-related system prompts could enable tailor-made jailbreak attacks. Selective transparency measures and controlled disclosures can potentially help enable a balancing of oversight and information sensitivity [197].

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 complex 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.

8 Statement of Contribution

This paper identifies a critical gap in AI safety and security evaluations: current approaches focus on FMs or deployed applications while missing components and their interactions. We introduce the *cascade* as an object of analysis. In doing so, we provide the foundation for further research directions examining interacting behaviours that current auditing approaches miss.

References

[1] Jennifer Cobbe, Michael Veale, and Jatinder Singh. Understanding accountability in algorithmic supply chains. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 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.

[2] David Gray Widder and Dawn Nafus. 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.

[3] Agathe Balayn, Yulu Pi, David Gray Widder, Kars Alfrink, Mireia Yurrita, Sohini Upadhyay, Naveena Karusala, Henrietta Lyons, Cagatay Turkay, Christelle Tessono, Blair Attard-Frost, and Ujwal Gadiraju. 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, page 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.

[4] Kornel Lewicki, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 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, pages 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.

[5] Jennifer Cobbe and Jatinder Singh. 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.

[6] Sherry Yang, Ofir Nachum, Yilun Du, Jason Wei, Pieter Abbeel, and Dale Schuurmans. Foundation Models for Decision Making: Problems, Methods, and Opportunities, March 2023. URL http://arxiv.org/abs/2303.04129. arXiv:2303.04129 [cs].

[7] Andrés Domínguez Hernández, Shyam Krishna, Antonella Maia Perini, Michael Katell, Sj Bennett, Ann Borda, Youmna Hashem, Semeli Hadjiloizou, Sabeehah Mahomed, Smera Jayadeva, Mhairi Aitken, and David Leslie. Mapping the individual, social and biospheric impacts of Foundation Models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 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.

[8] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Liu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu, and Lichao Sun. 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.

[9] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle

Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, July 2022. URL http://arxiv.org/abs/2108.07258. arXiv:2108.07258 [cs].

[10] Harini Suresh, Emily Tseng, Meg Young, Mary Gray, Emma Pierson, and Karen Levy. Participation in the age of foundation models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 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.

[11] Elisabeth Kirsten, Ivan Habernal, Vedant Nanda, and Muhammad Bilal Zafar. The impact of inference acceleration on bias of LLMs. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *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)*, pages 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/.

[12] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "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, page 1671–1685, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706363. doi: 10.1145/3658644.3670388. URL https://doi.org/10.1145/3658644.3670388.

[13] Nupoor Ranade, Marly Saravia, and Aditya Johri. Using rhetorical strategies to design prompts: a human-in-the-loop approach to make AI useful. *AI & SOCIETY*, 40(2):711–732, 2025.

[14] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 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.

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

[16] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When Large Language Models Meet Personalization, 2024. URL https://arxiv.org/abs/2304.11406.

[17] Hongru Cai, Yongqi Li, Wenjie Wang, Fengbin Zhu, Xiaoyu Shen, Wenjie Li, and Tat-Seng Chua. Large Language Models Empowered Personalized Web Agents. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 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.

[18] Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, and Aleksander Mądry. MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering, December 2024. URL http://arxiv.org/abs/2410.07095. arXiv:2410.07095 [cs].

[19] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as Agents, October 2023. URL http://arxiv.org/abs/2308.03688. arXiv:2308.03688 [cs].

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

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

[22] Rishi Bommasani, Dilara Soylu, Thomas I. Liao, Kathleen A. Creel, and Percy Liang. 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.

[23] Benjamin Laufer, Hamidah Oderinwale, and Jon Kleinberg. Anatomy of a Machine Learning Ecosystem: 2 Million Models on Hugging Face, 2025. URL https://arxiv.org/abs/2508.06811.

[24] Aspen Hopkins, Isabella Struckman, Kevin Klyman, and Susan S. Silbey. 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, page 209–227, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714825. doi: 10.1145/3715275.3732017. URL https://doi.org/10.1145/3715275.3732017.

[25] Agathe Balayn, Mireia Yurrita, Fanny Rancourt, Fabio Casati, and Ujwal Gadiraju. 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.

[26] Blair Attard-Frost and David Gray Widder. The ethics of ai value chains. *Big Data & Society*, 12(2):20539517251340603, 2025. doi: 10.1177/20539517251340603. URL https://doi.org/10.1177/20539517251340603.

[27] Sarah Huiyi Cen, Aspen Hopkins, Andrew Ilyas, Aleksander Madry, Isabella Struckman, and Luis Videgaray Caso. AI Supply Chains. Available at SSRN, April 2023. URL https://ssrn.com/abstract=4789403.

[28] Jatinder Singh, Jennifer Cobbe, and Chris Norval. Decision Provenance: Harnessing Data Flow for Accountable Systems. *IEEE Access*, 7:6562–6574, 2019. doi: 10.1109/ACCESS.2018.2887201.

[29] Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. The foundation model transparency index, 2023. URL https://arxiv.org/abs/2310.12941.

[30] Aspen Hopkins, Sarah H. Cen, Andrew Ilyas, Isabella Struckman, Luis Videgaray, and Aleksander Mądry. AI Supply Chains: An Emerging Ecosystem of AI Actors, Products, and Services, 2025. URL https://arxiv.org/abs/2504.20185.

[31] Ian Brown. Allocating accountability in AI supply chains. *Ada Lovelace Institute*, 2023.

[32] Arvind Narayanan and Sayash Kapoor. 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].

[33] Jatinder Singh, Ian Walden, Jon Crowcroft, and Jean Bacon. Responsibility & Machine Learning: Part of a Process. Available at SSRN, October 2016. URL https://ssrn.com/abstract=2860048.

[34] A. Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 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.

[35] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 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.

[36] Violet Turri, Katelyn Morrison, Katherine-Marie Robinson, Collin Abidi, Adam Perer, Jodi Forlizzi, and Rachel Dzombak. 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, page 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.

[37] Heike Felzmann, Eduard Fosch-Villaronga, Christoph Lutz, Aurelia Tamás-Montano, and Sebastian Ullrich. 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.

[38] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 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.

[39] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety, 2016. URL https://arxiv.org/abs/1606.06565.

[40] Gregory Falco, Ben Shneiderman, Julia Badger, Ryan Carrier, Anton Dahbura, David Danks, Martin Eling, Alwyn Goodloe, Jerry Gupta, Christopher Hart, Marina Jirotka, Henric Johnson, Cara LaPointe, Ashley J Llorens, Alan K Mackworth, Carsten Maple, Sigurður Emil Pálsson, Frank Pasquale, Alan Winfield, and Zee Kin Yeong. Governing AI safety through independent audits. *Nat. Mach. Intell.*, 3(7):566–571, July 2021.

[41] Richard Ren, Steven Basart, Adam Khoja, Alexander Pan, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Gabriel Mukobi, Ryan Hwang Kim, Stephen Fitz, and Dan Hendrycks. Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 68559–68594. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/7ebcdd0de471c027e67a11959c666d74-Paper-Datasets_and_Benchmarks_Track.pdf.

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

[43] Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 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.

[44] Woojeong Kim, Ashish Jagmohan, and Aditya Vempaty. SEAL: Suite for Evaluating API-use of LLMs, 2024. URL https://arxiv.org/abs/2409.15523.

[45] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. Black-Box Access is Insufficient for Rigorous AI Audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 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.

[46] Edward Kembery, Ben Bucknall, and Morgan Simpson. Position Paper: Model Access should be a Key Concern in AI Governance, 2024. URL https://arxiv.org/abs/2412.00836.

[47] Byung Do Chung, Sung Il Kim, and Jun Seop Lee. 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.

[48] Ahmad Ali Atieh Ali, Abdel-Aziz Ahmad Sharabati, Mahmoud Allahham, and Ahmad Yacoub Nasereddin. 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.

[49] Jize Wang, Zerun Ma, Yining Li, Songyang Zhang, Cailian Chen, Kai Chen, and Xinyi Le. GTA: A Benchmark for General Tool Agents, 2024. URL https://arxiv.org/abs/2407.08713.

[50] Erik Jones, Anca Dragan, Aditi Raghunathan, and Jacob Steinhardt. Automatically Auditing Large Language Models via Discrete Optimization. In *Proceedings of the 40th International Conference on Machine Learning*, pages 15307–15329. PMLR, July 2023. URL https://proceedings.mlr.press/v202/jones23a.html. ISSN: 2640-3498.

[51] Zhi-Yuan Chen, Shiqi Shen, Guangyao Shen, Gong Zhi, Xu Chen, and Yankai Lin. Towards Tool Use Alignment of Large Language Models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1382–1400, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.82. URL https://aclanthology.org/2024.emnlp-main.82/.

[52] Qiaoyu Tang, Ziliang Deng, Hongyu Lin, Xianpei Han, Qiao Liang, Boxi Cao, and Le Sun. ToolAlpaca: Generalized Tool Learning for Language Models with 3000 Simulated Cases, September 2023. URL http://arxiv.org/abs/2306.05301. arXiv:2306.05301 [cs].

[53] Ethem Alpaydin and Cenk Kaynak. Cascading classifiers. *Kybernetika*, 34(4):369–374, 1998.

[54] Kai Petersen, Claes Wohlin, and Dejan Baca. 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*, pages 386–400. Springer, 2009.

[55] Shreya Chappidi, Jennifer Cobbe, Chris Norval, Anjali Mazumder, and Jatinder Singh. Accountability Capture: How Record-Keeping to Support AI Transparency and Accountability (Re)shapes Algorithmic Oversight, 2025. URL https://arxiv.org/abs/2510.04609.

[56] Kaustubh Sridhar, Souradeep Dutta, Dinesh Jayaraman, and Insup Lee. REGENT: A Retrieval-Augmented Generalist Agent That Can Act In-Context in New Environments, 2025. URL https://arxiv.org/abs/2412.04759.

[57] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Xuanhe Zhou, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin

Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Guoliang Li, Zhiyuan Liu, and Maosong Sun. Tool Learning with Foundation Models. *ACM Comput. Surv.*, 57(4), December 2024. ISSN 0360-0300. doi: 10.1145/3704435. URL https://doi.org/10.1145/3704435.

[58] Cornelius Emde, Alasdair Paren, Preetham Arvind, Maxime Kayser, Tom Rainforth, Thomas Lukasiewicz, Bernard Ghanem, Philip H. S. Torr, and Adel Bibi. Shh, don't say that! Domain Certification in LLMs, 2025. URL https://arxiv.org/abs/2502.19320.

[59] Zekun Li, Baolin Peng, Pengcheng He, and Xifeng Yan. Evaluating the Instruction-Following Robustness of Large Language Models to Prompt Injection. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 557–568, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.33. URL https://aclanthology.org/2024.emnlp-main.33/.

[60] Anna Neumann, Elisabeth Kirsten, Muhammad Bilal Zafar, and Jatinder Singh. 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, page 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.

[61] Shaoteng Liu, Haoqi Yuan, Minda Hu, Yanwei Li, Yukang Chen, Shu Liu, Zongqing Lu, and Jiaya Jia. RL-GPT: Integrating Reinforcement Learning and Code-as-policy, 2024. URL https://arxiv.org/abs/2402.19299.

[62] Hyeong Kyu Choi, Xuefeng Du, and Yixuan Li. Safety-Aware Fine-Tuning of Large Language Models. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL https://openreview.net/forum?id=SqL94fLSM7.

[63] Saeid Asgari, Aliasghar Khani, and Amir Hosein Khasahmadi. 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.

[64] Shangzi Xue, Zhenya Huang, Jiayu Liu, Xin Lin, Yuting Ning, Binbin Jin, Xin Li, and Qi Liu. 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.

[65] Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A Survey on Mixture of Experts, August 2024. URL http://arxiv.org/abs/2407.06204. arXiv:2407.06204 [cs].

[66] Xinyu Li, Ruiyang Zhou, Zachary C. Lipton, and Liu Leqi. Personalized Language Modeling from Personalized Human Feedback, 2024. URL https://arxiv.org/abs/2402.05133.

[67] Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do LLMs Recognize Your Preferences? Evaluating Personalized Preference Following in LLMs, 2025. URL https://arxiv.org/abs/2502.09597.

[68] Yaaseen Mahomed, Charlie M. Crawford, Sanjana Gautam, Sorelle A. Friedler, and Danaë Metaxa. Auditing GPT's Content Moderation Guardrails: Can ChatGPT Write Your Favorite TV Show? In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 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.

[69] Yi Dong, Ronghui Mu, Yanghao Zhang, Siqi Sun, Tianle Zhang, Changshun Wu, Gaojie Jin, Yi Qi, Jinwei Hu, Jie Meng, Saddek Bensalem, and Xiaowei Huang. Safeguarding Large Language Models: A Survey, June 2024. URL http://arxiv.org/abs/2406.02622. arXiv:2406.02622 [cs].

[70] Lang Cao. 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.

[71] 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].

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

[73] 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].

[74] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, January 2023. URL http://arxiv.org/abs/2201.11903. arXiv:2201.11903 [cs].

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

[76] Pengyu Cheng, Yong Dai, Tianhao Hu, Han Xu, Zhisong Zhang, Lei Han, Nan Du, and Xiaolong Li. Self-playing Adversarial Language Game Enhances LLM Reasoning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 126515–126543. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e4be7e9867ef163563f4a5e90cec478f-Paper-Conference.pdf.

[77] Qiguang Chen, Libo Qin, Jiaqi Wang, Jinxuan Zhou, and Wanxiang Che. Unlocking the Capabilities of Thought: A Reasoning Boundary Framework to Quantify and Optimize Chain-of-Thought, 2024. URL https://arxiv.org/abs/2410.05695.

[78] Wenqi Zhang, Zhenglin Cheng, Yuanyu He, Mengna Wang, Yongliang Shen, Zeqi Tan, Guiyang Hou, Mingqian He, Yanna Ma, Weiming Lu, and Yueting Zhuang. Multimodal Self-Instruct: Synthetic Abstract Image and Visual Reasoning Instruction Using Language Model, 2024. URL https://arxiv.org/abs/2407.07053.

[79] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 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*, pages 6491–6501, Barcelona Spain, August 2024. ACM. ISBN 9798400704901. doi: 10.1145/3637528.3671470. URL https://dl.acm.org/doi/10.1145/3637528.3671470.

[80] Yuqi Zhu, Shuofei Qiao, Yixin Ou, Shumin Deng, Shiwei Lyu, Yue Shen, Lei Liang, Jinjie Gu, Huajun Chen, and Ningyu Zhang. KnowAgent: Knowledge-augmented planning for LLM-based agents. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 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/.

[81] Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. ∞Bench: Extending Long Context Evaluation Beyond 100K Tokens, February 2024. URL http://arxiv.org/abs/2402.13718. arXiv:2402.13718 [cs].

[82] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen

Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, et al. The Llama 3 Herd of Models, November 2024. URL http://arxiv.org/abs/2407.21783. arXiv:2407.21783 [cs].

[83] Zhiruo Wang, Zhoujun Cheng, Hao Zhu, Daniel Fried, and Graham Neubig. What Are Tools Anyway? A Survey from the Language Model Perspective, March 2024. URL http://arxiv.org/abs/2403.15452. arXiv:2403.15452 [cs].

[84] Yi Dong, Ronghui Mu, Yanghao Zhang, Siqi Sun, Tianle Zhang, Changshun Wu, Gaojie Jin, Yi Qi, Jinwei Hu, Jie Meng, Saddek Bensalem, and Xiaowei Huang. Safeguarding Large Language Models: A Survey, 2024. URL https://arxiv.org/abs/2406.02622.

[85] Ivoline Ngong, Swanand Kadhe, Hao Wang, Keerthiram Murugesan, Justin D. Weisz, Amit Dhurandhar, and Karthikeyan Natesan Ramamurthy. Protecting Users From Themselves: Safeguarding Contextual Privacy in Interactions with Conversational Agents, 2025. URL https://arxiv.org/abs/2502.18509.

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

[87] Helen Nissenbaum. 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.

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

[89] Alexandra Brintrup, George Baryannis, Ashutosh Tiwari, Svetan Ratchev, Giovanna Martinez-Arellano, and Jatinder Singh. Trustworthy, responsible, ethical AI in manufacturing and supply chains: synthesis and emerging research questions, May 2023. URL http://arxiv.org/abs/2305.11581. arXiv:2305.11581 [cs].

[90] Charles Perrow. *Normal accidents: Living with high risk technologies-Updated edition*. Princeton university press, 2011.

[91] Roel Dobbe. System Safety and Artificial Intelligence. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 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.

[92] Nathalie Diberardino, Clair Baleshta, and Luke Stark. Algorithmic Harms and Algorithmic Wrongs. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 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.

[93] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin Lee, Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, and Kim Hazelwood. Sustainable AI: Environmental Implications, Challenges and Opportunities. *Proceedings of Machine Learning and Systems*, 4:795–813, April 2022. URL https://proceedings.mlsys.org/paper_files/paper/2022/hash/462211f67c7d858f663355eff93b745e-Abstract.html.

[94] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from Language Models, December 2021. URL http://arxiv.org/abs/2112.04359. arXiv:2112.04359 [cs].

[95] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. Harms from Increasingly Agentic Algorithmic Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 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.

[96] Mélanie Gornet, Simon Delarue, Maria Boritchev, and Tiphaine Viard. Mapping AI ethics: a meso-scale analysis of its charters and manifestos. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 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.

[97] Daniel James Bogiatzis-Gibbons. Beyond Individual Accountability: (Re-)Asserting Democratic Control of AI. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 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.

[98] 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/.

[99] AI Safety Institute. 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].

[100] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. URL http://arxiv.org/abs/2307.09288. arXiv:2307.09288 [cs].

[101] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, October 2023. URL http://arxiv.org/abs/2310.06825. arXiv:2310.06825 [cs].

[102] Shijie Chen, Yu Zhang, and Qiang Yang. Multi-Task Learning in Natural Language Processing: An Overview. *ACM Comput. Surv.*, 56(12), July 2024. ISSN 0360-0300. doi: 10.1145/3663363. URL https://doi.org/10.1145/3663363.

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

[104] Alexandra Chouldechova, A. Feder Cooper, Abhinav Palia, Dan Vann, Chad Atalla, Hannah Washington, Emily Sheng, and Hanna Wallach. Red Teaming: Everything Everywhere All at Once. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL https://openreview.net/forum?id=KEggQCeDUA.

[105] Xiaojing Fan and Chunliang Tao. Towards Resilient and Efficient LLMs: A Comparative Study of Efficiency, Performance, and Adversarial Robustness, 2024. URL https://arxiv.org/abs/2408.04585.

[106] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic Evaluation of Language Models, October 2023. URL http://arxiv.org/abs/2211.09110. arXiv:2211.09110 [cs].

[107] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding, January 2021. URL http://arxiv.org/abs/2009.03300. arXiv:2009.03300 [cs].

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

[109] Yanzhao Qin, Tao Zhang, Tao Zhang, Yanjun Shen, Wenjing Luo, Haoze Sun, Yan Zhang, Yujing Qiao, Weipeng Chen, Zenan Zhou, Wentao Zhang, and Bin Cui. SysBench: Can Large Language Models Follow System Messages?, October 2024. URL http://arxiv.org/abs/2408.10943. arXiv:2408.10943 [cs].

[110] Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. 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-Datasets_and_Benchmarks.html.

[111] Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A StrongREJECT for Empty Jailbreaks, August 2024. URL http://arxiv.org/abs/2402.10260. arXiv:2402.10260 [cs].

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

[113] Dong Huang, Yuhao Qing, Weiyi Shang, Heming Cui, and Jie M. Zhang. EffiBench: Benchmarking the Efficiency of Automatically Generated Code, 2025. URL https://arxiv.org/abs/2402.02037.

[114] Yongliang Shen, Kaitao Song, Xu Tan, Wenqi Zhang, Kan Ren, Siyu Yuan, Weiming Lu, Dongsheng Li, and Yueting Zhuang. TaskBench: Benchmarking Large Language Models for Task Automation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 4540–4574. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/085185ea97db31ae6dcac7497616fd3e-Paper-Datasets_and_Benchmarks_Track.pdf.

[115] Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. Benchmarking LLMs via Uncertainty Quantification. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 15356–15385. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1bdcb065d40203a00bd39831153338bb-Paper-Datasets_and_Benchmarks_Track.pdf.

[116] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic Evaluation of Language Models, 2023. URL https://arxiv.org/abs/2211.09110.

[117] Hanjun Luo, Haoyu Huang, Ziye Deng, Xinfeng Li, Hewei Wang, Yingbin Jin, Yang Liu, Wenyuan Xu, and Zuozhu Liu. BIGbench: A Unified Benchmark for Evaluating Multi-dimensional Social Biases in Text-to-Image Models, 2025. URL https://arxiv.org/abs/2407.15240.

[118] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding, 2021. URL https://arxiv.org/abs/2009.03300.

[119] Kundan Krishna, Sanjana Ramprasad, Prakhar Gupta, Byron C Wallace, Zachary Chase Lipton, and Jeffrey P. Bigham. 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.

[120] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. BBQ: A Hand-Built Bias Benchmark for Question Answering, March 2022. URL http://arxiv.org/abs/2110.08193. arXiv:2110.08193 [cs].

[121] Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. Evaluating and Mitigating Discrimination in Language Model Decisions, December 2023. URL http://arxiv.org/abs/2312.03689. arXiv:2312.03689 [cs].

[122] Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs, January 2024. URL http://arxiv.org/abs/2311.04892. arXiv:2311.04892 [cs].

[123] Alejandro Salinas, Amit Haim, and Julian Nyarko. 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].

[124] Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 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].

[125] Ashwinee Panda, Xinyu Tang, Christopher A. Choquette-Choo, Milad Nasr, and Prateek Mittal. Privacy auditing of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=60Vd7QOXlM.

[126] Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Li Erran Li, Ruohan Zhang, Weiyu Liu, Percy Liang, Li Fei-Fei, Jiayuan Mao, and Jiajun Wu. Embodied Agent Interface: Benchmarking LLMs for Embodied Decision Making. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 100428–100534. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/b631da756d1573c24c9ba9c702fde5a9-Paper-Datasets_and_Benchmarks_Track.pdf.

[127] Benjamin Estermann, Luca A. Lanzendörfer, Yannick Niedermayr, and Roger Wattenhofer. PUZZLES: A Benchmark for Neural Algorithmic Reasoning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 127059–127098. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e5d1eaadeed651ba1021c09149db4b92-Paper-Datasets_and_Benchmarks_Track.pdf.

[128] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250/.

[129] Tara Safavi and Danai Koutra. Relational World Knowledge Representation in Contextual Language Models: A Review. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 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/.

[130] Evan Hernandez, Belinda Z. Li, and Jacob Andreas. Inspecting and editing knowledge representations in language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=ADtL6fgNRv.

[131] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-Following Evaluation for Large Language Models, November 2023. URL http://arxiv.org/abs/2311.07911. arXiv:2311.07911 [cs].

[132] Mantas Mazeika, Xuwang Yin, Rishub Tamirisa, Jaehyuk Lim, Bruce W. Lee, Richard Ren, Long Phan, Norman Mu, Adam Khoja, Oliver Zhang, and Dan Hendrycks. Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs, February 2025. URL http://arxiv.org/abs/2502.08640. arXiv:2502.08640 [cs].

[133] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment faking in large language models, December 2024. URL http://arxiv.org/abs/2412.14093. arXiv:2412.14093 [cs].

[134] Jiaming Ji, Kaile Wang, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Changye Li, Hantao Lou, Josef Dai, Yunhuai Liu, and Yaodong Yang. Language Models Resist Alignment: Evidence From Data Compression, 2024. URL https://arxiv.org/abs/2406.06144.

[135] Norman Mu, Jonathan Lu, Michael Lavery, and David Wagner. A Closer Look at System Message Robustness. October 2024. URL https://openreview.net/forum?id=YZqDyqYwFf.

[136] Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive Learning from Complex Explanation Traces of GPT-4, June 2023. URL http://arxiv.org/abs/2306.02707. arXiv:2306.02707 [cs].

[137] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 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].

[138] Qingyu Ren, Jie Zeng, Qianyu He, Jiaqing Liang, Yanghua Xiao, Weikang Zhou, Zeye Sun, and Fei Yu. Step-by-Step Mastery: Enhancing Soft Constraint Following Ability of Large Language Models, 2025. URL https://arxiv.org/abs/2501.04945.

[139] Desik Rengarajan, Sajad Mousavi, Ashwin Ramesh Babu, Vineet Gundecha, Avisek Naug, Sahand Ghorbanpour, Antonio Guillen, Ricardo Luna Gutierrez, and Soumyendu Sarkar. Imitation guided Automated Red Teaming. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL https://openreview.net/forum?id=fkOZjYwm2R.

[140] Zeming Wei, Yifei Wang, Ang Li, Yichuan Mo, and Yisen Wang. 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].

[141] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking Black Box Large Language Models in Twenty Queries, July 2024. URL http://arxiv.org/abs/2310.08419. arXiv:2310.08419 [cs].

[142] Haokun Lin, Haobo Xu, Yichen Wu, Jingzhi Cui, Yingtao Zhang, Linzhan Mou, Linqi Song, Zhenan Sun, and Ying Wei. Duquant: Distributing outliers via dual transformation makes stronger quantized llms. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 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.

[143] Vladimir Malinovskii, Denis Mazur, Ivan Ilin, Denis Kuznedelev, Konstantin Burlachenko, Kai Yi, Dan Alistarh, and Peter Richtarik. PV-Tuning: Beyond Straight-Through Estimation for Extreme LLM Compression, 2024. URL https://arxiv.org/abs/2405.14852.

[144] Shaona Ghosh, Prasoon Varshney, Makesh Narsimhan Sreedhar, Aishwarya Padmakumar, Traian Rebedea, Jibin Rajan Varghese, and Christopher Parisien. 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.

[145] Kaiqu Liang, Haimin Hu, Ryan Liu, Thomas L. Griffiths, and Jaime Fernández Fisac. RLHS: Mitigating misalignment in RLHF with hindsight simulation. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL https://openreview.net/forum?id=Io0ECcgFkh.

[146] Philip Amortila, Dylan J Foster, Nan Jiang, Akshay Krishnamurthy, and Zakaria Mhammedi. 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.

[147] Logan Barnhart, Reza Akbarian Bafghi, Stephen Becker, and Maziar Raissi. Aligning to What? Limits to RLHF Based Alignment, 2025. URL https://arxiv.org/abs/2503.09025.

[148] Logan Barnhart, Reza Akbarian Bafghi, Stephen Becker, and Maziar Raissi. Aligning to what? limits to RLHF based alignment. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7556–7591, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. URL https://aclanthology.org/2025.findings-naacl.421/.

[149] Margaret Mitchell, Avijit Ghosh, Alexandra Sasha Luccioni, and Giada Pistilli. Fully Autonomous AI Agents Should Not be Developed, 2025. URL https://arxiv.org/abs/2502.02649.

[150] Bhargav Kumar Konidena, Jesu Narkarunai Arasu Malaiyappan, and Anish Tadimarri. 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.

[151] Will Orr and Edward B. Kang. AI as a Sport: On the Competitive Epistemologies of Benchmarking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 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.

[152] Aisyah Putri and Minh Quang Tran. 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.

[153] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin Lee, Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, and Kim Hazelwood. Sustainable AI: Environmental Implications, Challenges and Opportunities. In D. Marculescu, Y. Chi, and C. Wu, editors, *Proceedings of Machine Learning and Systems*, volume 4, pages 795–813, 2022. URL https://proceedings.mlsys.org/paper_files/paper/2022/file/462211f67c7d858f663355eff93b745e-Paper.pdf.

[154] Sasha Luccioni, Yacine Jernite, and Emma Strubell. Power Hungry Processing: Watts Driving the Cost of AI Deployment? In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 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.

[155] Sabri Eyuboglu, Karan Goel, Arjun Desai, Lingjiao Chen, Mathew Monfort, Chris Ré, and James Zou. Model ChangeLists: Characterizing Updates to ML Models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 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.

[156] Agathe Balayn, Natasa Rikalo, Jie Yang, and Alessandro Bozzon. 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.

[157] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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, page 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.

[158] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, J. B. Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims, April 2020. URL http://arxiv.org/abs/2004.07213. arXiv:2004.07213 [cs].

[159] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment, 2024. URL https://arxiv.org/abs/2308.05374.

[160] Boming Xia, Qinghua Lu, Liming Zhu, Sung Une Lee, Yue Liu, and Zhenchang Xing. 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, page 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.

[161] Stuart McLennan, Amelia Fiske, Daniel Tigard, Ruth Muller, Sami Haddadin, and Alena Buyx. 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.

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

[163] Jennifer Cobbe and Jatinder Singh. Accounting for context in AI technologies. In *Handbook on Public Policy and Artificial Intelligence*, pages 94–108. Edward Elgar Publishing, June 2024.

[164] Anjalie Field, Amanda Coston, Nupoor Gandhi, Alexandra Chouldechova, Emily Putnam-Hornstein, David Steier, and Yulia Tsvetkov. 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, pages 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.

[165] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. "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, pages 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.

[166] Samuel Mayworm, Kendra Albert, and Oliver L. Haimson. 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*, pages 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.

[167] Inyoung Cheong, King Xia, K. J. Kevin Feng, Quan Ze Chen, and Amy X. Zhang. (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, page 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.

[168] David Gray Widder, Derrick Zhen, Laura Dabbish, and James Herbsleb. 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*, pages 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.

[169] Michael Madaio, Shivani Kapania, Rida Qadri, Ding Wang, Andrew Zaldivar, Remi Denton, and Lauren Wilcox. Learning about Responsible AI On-The-Job: Learning Pathways, Orientations, and Aspirations. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 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.

[170] Lixiang Yan, Vanessa Echeverria, Gloria Milena Fernandez-Nieto, Yueqiao Jin, Zachari Swiecki, Linxuan Zhao, Dragan Gašević, and Roberto Martinez-Maldonado. 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.

[171] Gaole He, Gianluca Demartini, and Ujwal Gadiraju. 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.

[172] Alexandra-Elena Gurita and Radu-Daniel Vatavu. 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.

[173] Anders Bruun, Niels van Berkel, Dimitrios Raptis, and Effie L-C Law. 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.

[174] Inha Cha and Richmond Y. Wong. 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.

[175] Michael Veale, Max Van Kleek, and Reuben Binns. 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, page 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.

[176] Sarah Sterz, Kevin Baum, Sebastian Biewer, Holger Hermanns, Anne Lauber-Rönsberg, Philip Meinel, and Markus Langer. On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 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.

[177] Chowon Kang, Yoonseo Choi, Yongjae Sohn, Hyunseung Lim, and Hwajung Hong. 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.

[178] Jamie Hancock, Ruoyun Hui, Jatinder Singh, and Anjali Mazumder. Trouble at Sea: Data and digital technology challenges for maritime human rights concerns. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 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.

[179] Ari Ball-Burack, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. Differential Tweetment: Mitigating Racial Dialect Bias in Harmful Tweet Detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 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.

[180] Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, and Jacquelyn Schneider. Escalation Risks from Language Models in Military and Diplomatic Decision-Making. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 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.

[181] Bianca Giulia Sarah Schor, Emma Kallina, Jatinder Singh, and Alan Blackwell. Meaningful Transparency for Clinicians: Operationalising HCXAI Research with Gynaecologists. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 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.

[182] Jianing Qiu, Kyle Lam, Guohao Li, Amish Acharya, Tien Yin Wong, Ara Darzi, Wu Yuan, and Eric J. Topol. 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.

[183] Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, Wuyue Wang, Yiqun Liu, and Minlie Huang. LegalAgentBench: Evaluating LLM agents in legal domain. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 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/.

[184] Mazda Moayeri, Elham Tabassi, and Soheil Feizi. Worldbench: Quantifying geographic disparities in llm factual recall. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 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.

[185] Dimitris Bertsimas and Yu Ma. M3H: Multimodal Multitask Machine Learning for Healthcare, 2024. URL https://arxiv.org/abs/2404.18975.

[186] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. Mdagents: An adaptive collaboration of llms for medical decision-making. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 79410–79452. Curran Associates, Inc., 2024. doi: 10.52202/079017-2522. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/90d1fc07f46e31387978b88e7e057a31-Paper-Conference.pdf.

[187] Anup Kumar. 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.

[188] Mina Valizadeh and Natalie Parde. The AI doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 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/.

[189] Youngduck Choi, Youngnam Lee, Dongmin Shin, Junghyun Cho, Seoyon Park, Seewoo Lee, Jineon Baek, Chan Bae, Byungsoo Kim, and Jaewe Heo. Ednet: A large-scale hierarchical dataset in education. In Ig Ibert Bittencourt, Mutlu Cukurova, Kasia Muldner, Rose Luckin, and Eva Millán, editors, *Artificial Intelligence in Education*, pages 69–73, Cham, 2020. Springer International Publishing. ISBN 978-3-030-52240-7.

[190] Cecilia Panigutti, Alan Perotti, André Panisson, Paolo Bajardi, and Dino Pedreschi. 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.

[191] Alexandra Brintrup, George Baryannis, Ashutosh Tiwari, Svetan Ratchev, Giovanna Martinez-Arellano, and Jatinder Singh. Trustworthy, responsible, ethical AI in manufacturing and supply chains: synthesis and emerging research questions, 2023. URL https://arxiv.org/abs/2305.11581.

[192] Konstantina Palla, José Luis Redondo García, Claudia Hauff, Francesco Fabbri, Henrik Lindström, Daniel R. Taber, Andreas Damianou, and Mounia Lalmas. 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].

[193] Beatrice Nolan. 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].

[194] Nadine Yousif. 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].

[195] N.G. Leveson. 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.

[196] Agathe Balayn, Lorenzo Corti, Fanny Rancourt, Fabio Casati, and Ujwal Gadiraju. Understanding Stakeholders' Perceptions and Needs Across the LLM Supply Chain, 2024. URL https://arxiv.org/abs/2405.16311.

[197] Chris Norval, Kristin Cornelius, Jennifer Cobbe, and Jatinder Singh. 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, page 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.