Position: Auditing Is Not Evaluating; LLM Audit Requires Dynamic, Contextual, Budget-Aware and Reliable Evidence

Clea Chataigner [1 2 3 4] **Pablo Piantanida** [2 3 4 5] **Golnoosh Farnadi** [1 4]

Abstract

Auditing large language models (LLMs) is increasingly urgent as these systems are deployed in high-stakes settings, yet existing evaluation practices are ill-suited to meet auditing requirements. Directly repurposing standard evaluation tools can yield incomplete or misleading conclusions, e.g. overstating robustness when evidence comes from static prompts rather than adaptive, real-world interactions. This position paper argues that effective LLM audits must instead generate dynamic, context-sensitive, budget-aware, and reliable evidence. To support this position, we analyze how each of these principles can be operationalized through a four-component framework: *Auditing Scope*, *Interactor*, *Evaluator*, and *Output*. We highlight design requirements, assumptions, limitations and research directions, demonstrating how high-level principles can be translated into concrete, actionable, evidence-based procedures.

1. Introduction

Artificial Intelligence (AI) auditing is becoming urgent as AI systems are deployed in high-stakes settings, from legal advice to medical diagnosis (Megan Ma, 2023; Esmaeilzadeh, 2024). Without credible oversight, such deployment can amplify misinformation, systemic bias, and other unintended harms (Suresh & Guttag, 2021; Chehbouni et al., 2024; Bengio et al., 2025). Consequently, regulatory initiatives such as the EU AI Act (Hupont et al., 2023) and the U.S. NIST AI Risk Management Framework (AI, 2023) require organizations to conduct audits, with phased compliance timelines beginning in the mid-2020s in the European Union.

Auditing has conventionally referred to an independent examination of complex processes, aimed at verifying compliance with standards or regulations (Gupta, 2004; IEEE, 2008; Mökander et al., 2024). Translating this logic to AI is difficult: classical audits presume relatively stable, well-specified processes, while modern AI systems are probabilistic, adaptive, and context-sensitive (Raji et al., 2020). These tensions are amplified for large language models (LLMs): they are general-purpose, interactive systems whose behavior shifts with prompts, users, and environments. Mature operational LLM auditing frameworks remain limited, with existing proposals articulating high-level principles (Mökander et al., 2024) but offering little practical guidance on how audits should be conducted or how audit evidence should be systematically collected.

Meanwhile, the technical literature repurposed to fill this gap—LLM evaluation—has evolved toward different goals. Many protocols emphasize static, benchmarked, single-turn settings (Hendrycks et al., 2021; Phan et al., 2025); rely on costly and inconsistent human judgments (Clark et al., 2021; Zhou et al., 2022; Howcroft et al., 2020); or use imperfect automated and LLM-based metrics (Fabbri et al., 2021; Bavaresco et al., 2025; Chehbouni et al., 2025). They also struggle to capture abstract, context-dependent behaviors of real deployments and often overstate what is actually measured (Wallach et al., 2025). Although recent work explores more dynamic and interactive evaluation (Yu et al., 2024; Zhu et al., 2023; Kim et al., 2025), most protocols remain centered on comparative performance rather than producing decision-relevant evidence aligned with auditing requirements, under realistic access and resource constraints (Blodgett et al., 2021; Wang et al., 2024b). For example, a model may pass static safety benchmarks yet fail in multi-turn conversations where user goals shift, conversation history accumulates, or tool use introduces new attack surfaces.

In this position paper, we argue that auditing is not evaluating: LLM audits must produce dynamic, contextual, budget-aware and reliable evidence. Evaluation estimates average performance under fixed conditions; auditing is a decision procedure that must justify whether a particular system meets a stated standard in a stated context, using a transparent protocol and a finite evidence budget (bounded queries, time, compute, and finite human review/labeling). This does not require testing all contexts: audits should

[1] McGill University, Montréal, Canada [2] Université Paris-Saclay, Paris, France [3] ILLS, Montréal, Canada [4] Mila, Montréal, Canada [5] CNRS, France. Correspondence to: Clea Chataigner <clea.chataigner@mila.quebec>.

Preprint. February 6, 2026.

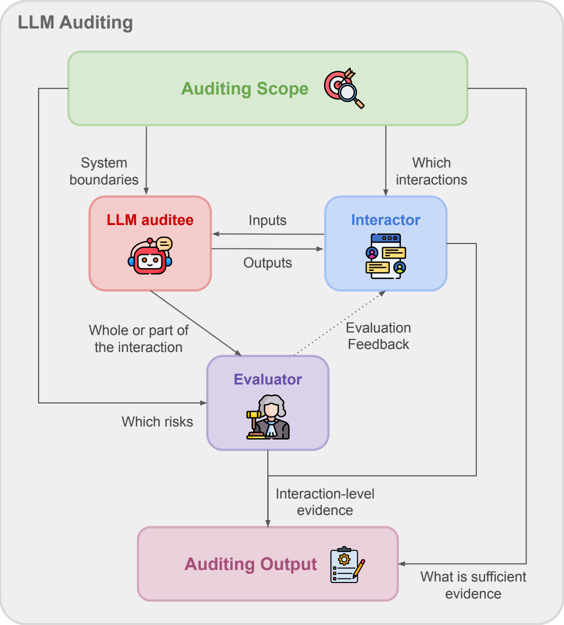

Figure 1. The process of LLM auditing.

declare scope and allocate budget to the highest-risk interaction patterns and user populations. As a result, auditing requires (i) interactions that reflect realistic use cases and user contexts, (ii) explicit accounting of constraints on time, computation, and human expertise, and (iii) tools adapted to yield interpretable, reliable, and actionable evidence rather than aggregate scores.

This paper is organized as follows. We first review related work on LLM auditing and evaluation in Section 2. We then introduce a four-component framework for LLM auditing in Section 3: the *Auditing Scope* (what behaviors and decisions are in scope, and what evidence matters), the *Interactor* (how the audit probes the system), the *Evaluator* (how outputs are interpreted and validated), and the *Auditing Output* (how evidence is compared to regulations and summarized into conclusions and guarantees) (Figure 1). Section 4 uses this framework to articulate key distinctions between auditing and traditional evaluation. Finally, Section 5 presents research directions, and Section 6 analyzes alternative views.

2. Related work

2.1. AI and LLM auditing

Auditing has traditionally been understood as an independent review of complex processes to evaluate whether they adhere to established standards or regulations (Gupta, 2004; IEEE, 2008; Mökander et al., 2024). Transferring these ideas to AI is non-trivial: classical audits assume stable, well-specified processes, whereas AI systems are adaptive, probabilistic, and often opaque (Raji et al., 2020).

This has motivated the emergence of AI auditing as a distinct, multidisciplinary field spanning computer science, law, organizational studies, and ethics (Sandvig et al., 2014; Kearns et al., 2018; Selbst, 2021; Bandy, 2021; Floridi et al., 2022; Minkkinen et al., 2022). Frameworks have been proposed across technical, legal, and organizational dimensions (Raji et al., 2020; Metaxa et al., 2021; Mökander et al., 2021; Brown et al., 2021), highlighting that risks often arise from system design, deployment context, and governance rather than model behavior alone.

For LLMs and other foundation models, however, concrete technical auditing procedures remain underdeveloped despite widespread deployment. Mökander et al. (2024) highlight this gap and argue that LLM audits require methodological adaptations, proposing a three-layer framework spanning governance, model, and application audits. While there is growing consensus on LLM auditing goals, there is limited guidance on operationalizing these goals into concrete, evidence-based procedures (Li & Goel, 2025; Costanza-Chock et al., 2022; Guha et al., 2024). Existing proposals largely stop at high-level principles and rely on limited technical tools, often centered on static benchmarks. Defining audit principles alone is insufficient without specifying what evidence is collected, under what constraints, and what claims it supports.

The goal of this paper is to help bridge this gap by focusing on the operationalization of technology audits for LLMs. We deliberately exclude governance and ecosystem audits (Mökander et al., 2024; Birhane et al., 2024), which raise distinct methodological and sociotechnical challenges. AI audits vary along several other dimensions (Table 1, Appendix A) but our primary focus is on *compliance-based audits*, assessing adherence to predefined standards, rather than *risk-based audits*, aimed at informing model development. We further restrict our analysis to *black-box auditing* and discuss these limitations in Section 7.

2.2. LLM evaluation

LLMs are general-purpose, generative, and highly adaptive systems. Consequently, the targets of their evaluation, i.e. capabilities, behaviors, or societal impacts, are often abstract, context-dependent, and entangled with human values and societal norms (Wallach et al., 2025), making them challenging to evaluate.

Historically, evaluation has addressed measurement challenges using closed-form generation tasks (e.g multiple-choice questions), where scoring is straightforward (Hendrycks et al., 2021). However, such tasks oversimplify real-world usage and fail to capture the complexity of open-ended generation (Wei et al., 2024). This has motivated a shift toward evaluating open-ended outputs, such as summaries, stories, or dialogue. Traditional approaches rely on human judgment along multiple dimensions, including factuality, coherence, or creativity (Fabbri et al., 2021). While human evaluation remains the gold standard, it is costly, difficult to scale, and prone to inconsistency (Clark et al., 2021; Zhou et al., 2022; Howcroft et al., 2020).

Automated metrics offer scalable alternatives. ROUGE (Lin, 2004) measures n-gram overlap and evaluates surface-level similarity and content coverage, while perplexity measures the predictive likelihood of a model as a proxy for fluency. BERTScore (Zhang et al., 2020) leverages contextual embeddings to assess semantic similarity between generated and reference text. Despite their efficiency, these metrics correlate poorly with human judgment for complex, open-ended generation tasks (Fabbri et al., 2021).

These limitations have contributed to the widespread adoption of LLMs-as-Judges (LLJs). LLJs use LLMs to automatically assess generated outputs along diverse dimensions, such as factuality, coherence, or ethical considerations, enabling more nuanced and scalable evaluation than traditional automated metrics (Liu et al., 2023). However, LLJs introduce new challenges regarding validity, reliability, and consistency, raising questions about whether their judgments truly reflect human preferences (Bavaresco et al., 2025; Rawte et al., 2023; Chehbouni et al., 2025).

While these developments address how outputs are evaluated, most evaluation protocols still rely on static, benchmark-centered inputs and where performance is evaluated in one isolated input–output interaction (single-turn) (Phan et al., 2025). Such settings are well suited for estimating average performance under controlled conditions, but they seriously differ from realistic interaction patterns (Blodgett et al., 2021; Wang et al., 2024b). Recent work has begun to explore more dynamic and interactive evaluation settings, including LLM-based interactors that adapt prompts based on model responses (Yu et al., 2024; Zhu et al., 2023; Kim et al., 2025), but these approaches bring their own limitations when repurposed for auditing, which we discuss in detail below.

Overall, current LLM evaluation remains focused on comparative performance, offering little guidance on how to collect, prioritize, or interpret evidence under realistic constraints. This motivates our core claim: existing evaluation tools do not directly support auditing objectives.

3. A Unified Framework for LLM Auditing

Building on gaps identified in prior work (Li & Goel, 2025; Costanza-Chock et al., 2022; Guha et al., 2024), we now address the practical question of how to operationalize LLM audits. Existing work often treats auditing objectives, tools, and evaluation methods in isolation, leaving the overall process unclear. To address this, we propose a unified framework decomposing LLM auditing into core components.

The framework builds on two literatures: *Auditing Scope* and *Auditing Output* come from the AI auditing literature, which emphasizes evidentiary standards and reporting requirements (Raji et al., 2020; Cen & Alur, 2024); *Interactor* and *Evaluator* are inspired by LLM evaluation research, including dynamic and adaptive testing methods (Yu et al., 2024; Zhu et al., 2023; Kim et al., 2025). By bridging these perspectives, the framework clarifies how evidence is generated, interpreted, and communicated, supporting systematic audit design.

Before introducing the four components, we define the *LLM Auditee* as the system under audit. This may be an isolated LLM, an LLM augmented with safeguards or monitoring mechanisms (Neumann & Singh, 2025), or a multi-agent system. For simplicity, we focus on a single LLM, whether or not it is embedded in a larger system, and defer discussion of multi-agent audits to Section 7. We then formalize LLM auditing as a four-component process (Figure 1):

Auditing Scope The auditee is situated within an explicit *Auditing Scope*, which specifies the task or capability under examination, the boundaries of the system under audit (e.g., whether to include the system prompt, guardrails, monitoring mechanisms, user interfaces, and deployment context), the operational constraints, such as the total number of queries, prompts, or API calls permitted during the audit, and the stopping criteria. The scope also determines the evidence needed to certify that the auditee meets predefined risk, performance, or compliance criteria.

Interactor The *Interactor* is the component responsible for interacting with the LLM auditee. Interactions may be generated by humans, LLMs, or a combination of both. They can be (i) static, where all prompts are fixed in advance (e.g., benchmarks); (ii) partially dynamic, where initial prompts are fixed but follow-up questions adapt to model responses; or (iii) fully dynamic, where interactions evolve entirely online. The interactor's objectives may also vary. Interactions can be adversarial, aiming to elicit failures; collaborative, providing feedback or guidance; or observative, intended to probe behavior without steering the model.

Evaluator The *Evaluator* component assesses the outputs produced during interactions between the *Interactor* and the *LLM auditee*. Evaluators may take several forms: automated metrics, human judgment, LLJs, or a combination. The evaluator may operate at different levels, scoring individual responses or the whole interaction. In some settings, evaluation outcomes may further guide the *Interactor*, enabling adaptive auditing procedures.

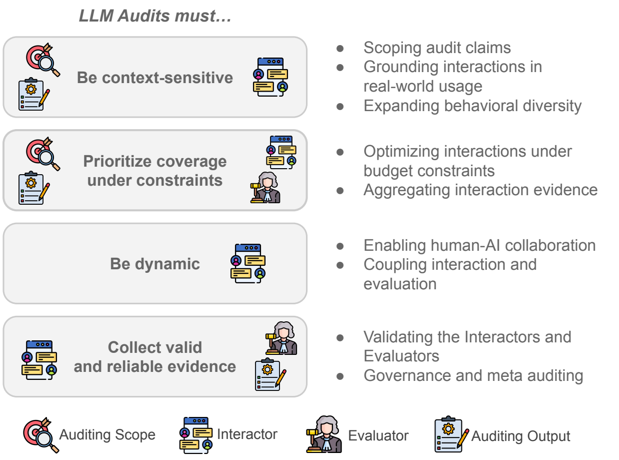
LLM Audits must...
Be context-sensitive
Scoping audit claims
Grounding interactions in real-world usage
Expanding behavioral diversity
Prioritize coverage under constraints
Optimizing interactions under budget constraints
Aggregating interaction evidence
Be dynamic
Enabling human-AI collaboration
Coupling interaction and evaluation
Collect valid and reliable evidence
Validating the Interactors and Evaluators
Governance and meta auditing
Auditing Scope
Interactor
Evaluator
Auditing Output

Figure 2. Key operational principles for LLM audits.

Auditing Output The audit produces an explicit *Auditing Output*, presenting evidence and specifying claims or guarantees within the defined scope. Here, evidence refers to the collected observations, measurements, or artifacts, such as model responses, interaction logs, failure cases, or metric scores, that support conclusions about the auditee's compliance. Outputs may include quantitative summaries, structured reports, annotated examples of failures, or formal compliance statements. Unlike evaluation scores, auditing outputs are context-bound, interpretable, traceable, and explicit about limitations, stating what was examined, under which conditions, and with what confidence.

4. Operational Principles of LLM Auditing

Having decomposed LLM auditing into four core components, we articulate four operational principles that distinguish auditing from evaluation. Audits must be context-sensitive, cover relevant behaviors under resource constraints, support dynamic interactions, and collect valid and reliable evidence (Figure 2). Each principle highlights common evaluation-derived failure modes and motivates requirements on specific components of our framework.

We illustrate each principle using three auditing use cases: (i) LLMs for code generation used in professional and educational settings (Wang et al., 2025b); (ii) LLMs used for legal research and practice (Megan Ma, 2023); and (iii) educational LLM tools that provide writing and application support to students (Prabhudesai et al., 2025).

4.1. Audits must be context-sensitive

We begin by arguing that LLM audits are inherently **context-sensitive**: audit claims are only meaningful when grounded in the specific tasks, risks, and usage settings of a system.

By contrast, general-purpose evaluation often measures abstract capabilities without grounding them in real-world contexts (Hendrycks et al., 2021; Phan et al., 2025), making it unclear what is being measured and why (Wallach et al., 2025). Audits, however, operate differently: they test contextual hypotheses rather than global properties. This requires auditors to specify what evidence would suffice to reject a presumption of compliance within a clearly defined scope (Cen & Alur, 2024). Defining that *Auditing Scope*, with tasks, domains, usage contexts, audit targets, and advocacy objectives, is essential to produce interpretable, actionable, and relevant evidence (Birhane et al., 2024; Raji & Buolamwini, 2019). It also clarifies the limits of audit claims in the *Auditing Output*: a system may perform acceptably in one context while failing in another.

Crucially, these contextual specifications directly shape *Interactor* design by determining which interactions are permissible, meaningful, and representative. Yet interactors derived from evaluation settings often struggle to meet these requirements. Benchmark-based interactors risk producing artificial and decontextualized interactions (Blodgett et al., 2021; Crockett & Messeri, 2025), while human interactors may overlook critical behaviors if they do not reflect the target user population (DeVos et al., 2022; Gadiraju et al., 2023; Kapania et al., 2023; Fan et al., 2022). LLM-based interactors introduce complementary challenges: they can generate unnatural prompts (Jones et al., 2023; Lee et al., 2025a) and exhibit output homogenization (Wu et al., 2025; Sourati et al., 2025), including mode and knowledge collapse (Jiang et al., 2025; Zhang et al., 2025a;b; Wright et al., 2025). These effects can systematically exclude certain demographics or usage patterns (Chataigner et al., 2025; Agnew et al., 2024).

These risks appear across our use cases, underscoring the need to clearly define the *Auditing Scope*. For coding LLMs, should the *Interactor* mimic expert developers or beginner students? Legal LLMs may require fact-checking mechanisms to detect fabricated cases or counterfactual hallucinations (Weiser, 2023; Dahl et al., 2024) while conversational LLMs demand attention to stereotypes, cultural norms, and other biases. Context also shapes the audit process: testing educational LLMs with hiring scenarios outside the system's deployment (Prabhudesai et al., 2025) produces unrealistic behaviors and offers no real-world insight. Similarly, auditing legal LLMs without accounting for jurisdiction-specific terminology can elicit fluent but irrelevant responses.

4.2. Audits must prioritize coverage under resource constraints

Even when relevant tasks, risks, and usage settings are clearly specified, auditors must still decide how extensively these contexts are explored under finite resources. This motivates a second operational principle: **coverage under resource constraints**.

Auditing faces a fundamental trade-off between coverage and budget. Meaningful audits require structured coverage, ensuring that relevant usage contexts, stakeholder groups, and plausible failure modes are examined (Sandvig et al., 2014; Raji et al., 2020; Birhane et al., 2024; Mökander et al., 2024), while operating under limited queries, time, compute, and expert review. The auditing process must therefore include a principled stopping rationale under an explicit budget, specified in the *Auditing Scope* and reflected in the design of both the *Interactor* and the *Evaluator*.

By contrast, evaluation protocols typically rely on implicit or arbitrary stopping criteria. Benchmarks fix the number of test items, treat them as equally informative and assume that aggregated performance sufficiently characterizes the system. However, benchmark items vary widely in difficulty and diagnostic value (Hofmann et al., 2025; Siska et al., 2024), and uniform sampling can obscure rare but high-risk behaviors. Under resource constraints, not all interactions should be treated equally: the *Interactor* should prioritize interactions by expected information value, and these decisions should be documented in the *Auditing Output*, explaining why certain evidence was collected and others were not (Cen & Alur, 2024).

We can highlight such trade-offs through our use cases. In coding audits, simple syntactic fixes are inexpensive to test but provide limited diagnostic value, while multi-file debugging or iterative refactoring are more informative yet costly (Jain et al., 2025). In legal auditing, scarce expert fact-checking forces trade-offs between broad domain coverage and deeper jurisdiction-specific analysis. In educational settings, auditors cannot cover all student populations, requiring explicit assumptions about demographics and trade-offs between many short interactions and fewer, longer pedagogical exchanges (Wang et al., 2024a).

4.3. Audits must be dynamic

Building on the need for context-sensitive coverage under resource constraints, we argue that static interactions alone cannot satisfy these requirements, making **dynamic interaction** essential.

First, real users interact with LLMs through open-ended, multi-turn exchanges (Zhao et al., 2024), yet static benchmarks are inherently single-turn and most evaluation protocols neglect this interactive dimension (Wang et al., 2024b; Pan et al., 2025). As discussed in Section 4.1, audits should reflect actual user behavior; a static *Interactor* therefore fails to capture behaviors that emerge only over sustained interaction. In our first use case, coding with a LLM is highly collaborative: users refine code across multiple turns, test outputs, and debug errors. Many failures such as compounding bugs surface only through this back-and-forth and are invisible to single-turn prompts (Pan et al., 2025).

Second, static benchmarks are vulnerable to contamination and memorization. When benchmark items appear in training data, auditee models may retrieve memorized answers rather than engage in genuine reasoning, inflating apparent performance and masking real weaknesses (Alzahrani et al., 2024; Zhou et al., 2023). In our second use case, directly reusing legal benchmarks such as LawBench (Fei et al., 2024) or LegalBench (Guha et al., 2023), which may overlap with training data, can overstate faithfulness while failing to surface real legal hallucinations. Dynamic interaction lets the *Interactor* explore novel inputs, producing evidence that better reflects deployment behavior (Kim et al., 2025; Li et al., 2025a; Yu et al., 2024).

Third, auditing under explicit budget constraints requires adaptivity. As examined in Section 4.2, interactions should be prioritized to spend resources efficiently. While static interaction treats all queries as independent, adaptive interaction allows the *Interactor* to condition future queries on observed outputs (Kim et al., 2025; Yu et al., 2024; Bai et al., 2023). When coupled with intermediate signals from the *Evaluator*, this helps concentrate effort where risk or uncertainty appears highest, for example, pushing further along a failing line of reasoning in a coding audit, ensuring that limited resources are used efficiently and diagnostically.

4.4. Audits must collect valid and reliable evidence

Even when the preceding principles are met, the **reliability and validity** of interactions and evaluations remain critical: flawed *Interactor* and *Evaluator* components can compromise the collected evidence, and in turn, undermine the conclusions reported in the *Auditing Output*.

Dynamic evaluation typically assumes that follow-up interactions are coherent, relevant, and informative (Li et al., 2025a; Yu et al., 2024; Bai et al., 2023). However, this assumption is fragile. LLM-based interactors may hallucinate, produce incoherent steps, or reinforce prior errors rather than reveal new failures (Laban et al., 2025; Gorle et al., 2025). Some frameworks justify the use of LLM-based interactors by citing their effectiveness in synthesis and reasoning tasks (Amirizaniani et al., 2025). This presupposes that reasoning competence translates into reliable interactions, an assumption that remains largely unvalidated. Human interactors also face limitations: without domain expertise, familiarity with realistic usage, or auditing experience, follow-ups may fail to elicit meaningful behaviors. For example, in educational settings, student auditors struggled to conceptualize and evaluate LLM biases on their own (Prabhudesai et al., 2025).

Even if interactions are well-designed, evaluations themselves are not inherently reliable. LLJs are often treated as proxies for human judgment (Chehbouni et al., 2025), but this assumes well-defined constructs and correct annotations, which are rarely scrutinized. LLJs also exhibit documented limitations in instruction adherence, robustness, domain expertise, and faithfulness (Chehbouni et al., 2025; Hu et al., 2024; Agarwal et al., 2024; Si et al., 2024). A particular concern is self-enhancement bias (Liu et al., 2024a), where models favor outputs from the same or closely related model families, a risk amplified when the auditee might incorporate an unknown LLM from the same family as the *Evaluator*. Human evaluation is also imperfect: judgments are influenced by task framing, incentives, and annotator background, leading to noise and inconsistency (Clark et al., 2021; Howcroft et al., 2020; Zhou et al., 2022).

Importantly, in a dynamic process, the *Evaluator* can provide intermediate signals to the *Interactor* guiding future queries or triggering early stopping. If the evaluator is flawed or miscalibrated, it can misguide the interactions (Shen et al., 2025), leading to incomplete or biased evidence. For example, Yu et al. (2024) allow the evaluator to terminate interactions early when a response is deemed "egregiously inadequate", which can prematurely halt exploration and miss informative failures. In coding audits, such errors can cause the *Interactor* to focus on false mistakes of the auditee instead of exploring other behaviors.

In summary, effective LLM auditing requires four interdependent principles: grounding audits in real-world contexts, ensuring coverage under explicit resource constraints, supporting dynamic interactions, and collecting valid, reliable evidence. These principles illustrate why traditional evaluation tools are insufficient for auditing: they are often static, decontextualized, and prone to unreliable measurement.

5. Research directions

The principles above highlight open challenges for LLM auditing that current evaluation and auditing practices do not resolve. This section outlines research directions arising from these challenges, focusing on formalizing audit claims, designing interactions under resource constraints, and collecting valid, reliable evidence.

5.1. Contextualization and audit scope

A central research challenge for LLM auditing is moving beyond decontextualized evaluations toward making claims that are explicitly scoped, interpretable, and decision-relevant.

Scoping audit claims. A central research direction is clarifying what audit evidence can legitimately support about the *LLM auditee*. Building on work on validity and reliability (Salaudeen et al., 2025; Weidinger et al., 2025; Wallach et al., 2025), future research should formalize how different forms of evidence support different kinds of claims, and how confidence in those claims depends on the quantity, diversity, and quality of audited interactions. Making these limits explicit ensures that audit findings are interpreted appropriately and used effectively for governance and decision-making.

Grounding interactions in real-world usage. Research should design interaction-generation pipelines that reflect actual user behavior rather than artificial prompts. This can include drawing on real user queries (Zhao et al., 2024; Aerni et al., 2024; Jiang et al., 2025) or user-reported errors and failures (Deng et al., 2025; Cabrera et al., 2021). Automated generation can be guided by linguistic or structural constraints to better match realistic usage contexts (Chataigner et al., 2025; Lee et al., 2025b). Persona-based generation may further improve realism (Abdulhai et al., 2025), though care is needed, as simplified personas can misrepresent real users (Wang et al., 2025a).

Expanding behavioral diversity. Improving behavioral coverage requires actively countering output homogenization. Relevant techniques include verbalized or diversity-aware sampling (Zhang et al., 2025a) and varying decoding strategies for the *Interactor*: higher-temperature settings could generate more varied prompts (Li et al., 2025a; Troshin et al., 2025), although some work favor deterministic decoding for reproducibility (Yu et al., 2024; Bai et al., 2023). The extent to which these methods capture audit-relevant diversity remains an open question (Peeperkorn et al., 2024). For human interactors, diversity-aware annotation and recruitment strategies provide complementary mechanisms to ensure coverage across user groups and usage patterns (Parrish et al., 2024).

5.2. Coverage, adaptivity, and budget-aware auditing

Auditing LLMs requires maximizing coverage of relevant behaviors while staying within limited budgets. Adaptive strategies and principled aggregation ensure that interactions are informative, and that reported evidence clearly reflects both what was explored and what remains unexamined.

Optimizing interactions under budget constraints. Auditing under limited budgets requires principled prompt selection. Adaptive strategies help focus resources on the most informative interactions, including Item Response Theory to estimate prompts likely to reveal meaningful behavior (Hofmann et al., 2025), targeted stress testing through automated red-teaming (Jones et al., 2023), structured exploration with tree-based strategies (Li et al., 2025a), and Information Theory metrics to quantify information gain and redundancy (Gorle et al., 2025). Uncertainty estimates can further guide early stopping (Farquhar et al., 2024; Kuhn et al., 2023; Yadkori et al., 2024; Li et al., 2025b). Such methods need to be reported in the *Auditing Output* to justify why certain evidence was collected and others were not, and why auditing stopped where it did.

Aggregating interaction-level evidence. A key direction for the *Auditing Output* is moving beyond simple score average to more meaningful aggregation. Future work could model the diagnostic value of interactions, for example by weighting evidence based on difficulty, rarity, or risk relevance (Hofmann et al., 2025; Li et al., 2025a; Truong et al., 2025). Besides, beyond providing quantitative results, the *Auditing Output* needs to incorporate qualitative analysis (Birhane et al., 2024). For such tasks, LLMs may assist in synthesizing interactions into higher-level summaries (Kim et al., 2025) or selecting the most interesting interactions as examples. However, the relevance and completeness of such tools must be carefully scrutinized, as poor synthesis or selective emphasis could undermine the audit's conclusions.

5.3. Validity, reliability, and governance of audit evidence

Auditing outcomes are only useful if both interactions and evaluations are valid, reliable, and properly governed. Achieving this requires coordinated human–AI collaboration, explicit interfaces between interaction and evaluation, and systematic validation of both processes.

Enabling human–AI collaboration. Combining the complementary strengths of LLMs and humans can improve both efficiency and reliability of the *Interactor*. LLMs can rapidly generate candidate interactions, many of which may be invalid, while humans can efficiently judge test validity but generate new tests more slowly and with high variability (Ribeiro & Lundberg, 2022; Rastogi et al., 2023). Similarly, the *Evaluator* can combine scalable LLJs with targeted human oversight, allocating routine or low-risk judgments to LLMs and reserving human expertise for domain-specific, high-impact, or uncertain cases (Pan et al., 2024; Sterz et al., 2024; Ashktorab et al., 2025). Open research questions include how to route cases between evaluators, integrate human and model feedback, and constrain these workflows under a budget (Shankar et al., 2024; Miranda et al., 2025).

Coupling interaction and evaluation. To support adaptive auditing, interaction and evaluation must be tightly coupled. The *Evaluator* should be able to defer judgment, request additional evidence, or signal uncertainty. Abstention mechanisms, such as those explored by Li et al. (2024), provide a foundation for deciding when evaluation should proceed versus when further interaction is needed. Future research should explore how making this interface explicit supports more principled stopping criteria and ensures that evidence collection and judgment remain tightly coupled.

Validating the interactors and evaluators. All auditing tools require systematic validation. An LLM *Interactor* should be evaluated for coverage, consistency, and adherence to realistic usage patterns, using human annotation for instance, rather than assumed reliable by default (Li et al., 2025a; Yu et al., 2024; Bai et al., 2023). Similarly, validating LLJs should go beyond simple correlation with human judgments, for example by aligning the entire judgment distribution with human responses (Chen et al., 2025), quantifying model–human agreement (Polo et al., 2025), and modeling rating indeterminacy when multiple interpretations are plausible (Guerdan et al., 2025). Techniques such as confidence scoring and uncertainty estimation can further flag inconsistent or low-reliability judgments (Farquhar et al., 2024; Kuhn et al., 2023; Yadkori et al., 2024; Li et al., 2025b). These approaches help surface ambiguous cases, contested norms, or edge behaviors, enabling audits to produce robust, evidence-based claims without collapsing diverse evaluative perspectives into a single score. Improving human *Interactor* and *Evaluator* is equally critical, for instance by improving experimental design and performing statistical evaluations (Schuff et al., 2023).

Governance and meta-auditing. Finally, auditing outputs themselves must be open to scrutiny. Questions of transparency, reproducibility, and accountability extend to the audit process and its conclusions. Prior work on participatory accountability and meta-auditing asks who defines auditing criteria, who evaluates audit quality, and whose interests audits ultimately serve (Costanza-Chock et al., 2022). Developing standards, documentation practices, and external review mechanisms is a critical direction for ensuring that auditing outputs are trustworthy and actionable.

6. Alternative views

In this section, we present several perspectives that provide alternatives to our position and discuss their merits and limitations.

Evaluation is equivalent to auditing. Some may argue that LLM auditing is simply a specialized form of evaluation, and that improved evaluation tools could be directly repurposed for auditing. Indeed, many challenges we raise, such as validity and reliability of LLJs, are well-known in the LLM evaluation literature, and in principle could be addressed within that community (Salaudeen et al., 2025; Weidinger et al., 2025; Wallach et al., 2025). LLM evaluation and auditing also share conceptual elements: both operate under finite budgets, and define tasks or contexts that guide measurement.

However, auditing imposes distinct requirements that standard evaluation tools do not fully satisfy. Audits differ in both scope and output: evaluation often summarizes average performance or informs model development, whereas auditing produces evidence-based guarantees, qualitative insights, and actionable claims about system behavior. Consequently, while LLM auditing necessarily relies on evaluation tools, these tools must be designed, fine-tuned, or optimized with explicit auditing objectives in mind, rather than assuming they can be directly repurposed.

Audits should be based on fixed, undisclosed benchmarks. Others argue that LLM auditing should rely on fixed, undisclosed benchmarks. Such tools are indeed reproducible, easier to validate, and, when undisclosed, robust to data contamination. Existing LLM auditing frameworks already incorporate standardized benchmarks (e.g. GLUE, SuperGLUE, BIG-bench) (Mökander et al., 2024), and static audits have demonstrated real-world impact such as Gender Shades for bias in facial analysis (Buolamwini & Gebru, 2018; Raji & Buolamwini, 2019). Moreover, some high-stakes applications like resume screening do not require multi-turn interaction, suggesting auditing research should focus on improving benchmark design rather than developing dynamic audits (Liu et al., 2024b).

Yet static audits cannot capture out-of-distribution behavior or support adaptive exploration under budget constraints (Lunardi et al., 2025; Cohen-Inger et al., 2025; Kim et al., 2025; Yu et al., 2024; Bai et al., 2023). Effective auditing requires conditioning future probes on observed outputs to prioritize areas of high uncertainty or potential harm. Purely static audits may suffice for narrowly scoped settings but fail to generalize to most real-world LLM deployments.

Audits should only be conducted by humans. Finally, some may argue that reliable audit evidence requires exclusively expert human judgment, especially in high-stakes domains such as election-related queries (Palta et al., 2024) or legal reasoning (Cheong et al., 2024), while we discussed several automated components in this paper. Human interactors and evaluators can capture subtle, context-specific issues that automated systems often miss, and across auditing practice, humans remain central due to their ability to interpret nuance, exercise professional judgment, and contextualize findings (Sterz et al., 2024). Human judgment is also widely regarded as the gold standard in LLM evaluation, with automated metrics typically validated against it (Fabbri et al., 2021).

This paper does not argue for replacing or broadly augmenting human auditors with AI. Rather, we focus on structuring LLM audits to produce valid, reliable evidence. Within this framing, relying solely on humans poses practical challenges: auditing is costly, time-consuming, and difficult to scale. Carefully designed automated components may therefore support specific parts of the process, such as interaction generation or preliminary evaluation, by increasing coverage or consistency. However, humans remain responsible for interpreting evidence, resolving ambiguity, and making audit claims. As emphasized throughout this paper, fully automated auditing carries substantial risks, including misinterpretation and output homogenization, making human oversight indispensable.

7. Limitations and Future Work

While our analysis clarifies key principles for single-LMM auditing, several limitations and open questions remain.

We concentrate on single LLMs, whereas an increasing body of work considers multi-agent or multi-model systems. While the framework we propose could be extended to such settings, for instance by placing the *Interactor* in direct interaction with multiple agents, this introduces additional challenges. Multi-agent audits require assumptions about the *Interactor*'s ability to coordinate, attribute, and adapt interactions across multiple LLMs; we leave these questions to future work.

We also focus on black-box access, which dominates practice due to IP constraints and API deployment. In contrast, white-box access enables techniques such as gradient-based attacks, mechanistic interpretability, or data-centric analyses (Casper et al., 2024). White-box signals like internal representations, confidence estimates, or training artifacts could inform prioritization, coverage, and uncertainty estimation, but integrating them into the *Interactor* and *Evaluator* would require substantially different assumptions and mechanisms than considered here.

More broadly, while this paper highlights key limitations of evaluation-derived tools when repurposed for auditing, additional failure modes and practical constraints will likely emerge as LLM auditing practices mature and diversify.

8. Conclusion

In this paper, we showed that LLM audits are decision-oriented procedures designed to justify claims about a specific system in a specific context, under explicit constraints on queries, time, computation, and human review. We argued that effective auditing rely on four interdependent principles, producing dynamic, contextual, budget-aware and reliable evidence, rather than relying on benchmark- or score-centric evaluations. By decomposing the process into four components (*Auditing Scope*, *Interactor*, *Evaluator*, and *Auditing Output*), we illustrated how these principles can be operationalized into concrete procedures and highlighted where current evaluation practices fall short. We hope this helps shift the focus from evaluating isolated performance metrics to designing audits that provide clear, reliable, and actionable insights.

References

Abdulhai, M., Cheng, R., Clay, D., Althoff, T., Levine, S., and Jaques, N. Consistently Simulating Human Personas with Multi-Turn Reinforcement Learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, October 2025. URL https://openreview.net/forum?id=A0T3piHiis.

Aerni, M., Rando, J., Debenedetti, E., Carlini, N., Ippolito, D., and Tramèr, F. Measuring Non-Adversarial Reproduction of Training Data in Large Language Models. In *The Thirteenth International Conference on Learning Representations*, October 2024. URL https://openreview.net/forum?id=590yfqz1LE.

Agarwal, C., Tanneru, S. H., and Lakkaraju, H. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models, 2024. URL https://arxiv.org/abs/2402.04614.

Agnew, W., Bergman, A. S., Chien, J., Díaz, M., El-Sayed, S., Pittman, J., Mohamed, S., and McKee, K. R. The Illusion of Artificial Inclusion. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 1–12, New York, NY, USA, May 2024. Association for Computing Machinery. ISBN 979-8-4007-0330-0. doi: 10.1145/3613904.3642703. URL https://dl.acm.org/doi/10.1145/3613904.3642703.

AI, N. Artificial intelligence risk management framework (ai rmf 1.0). *URL: https://nvlpubs. nist. gov/nistpubs/ai/nist. ai*, pp. 100–1, 2023.

Alzahrani, N., Alyahya, H., Alnumay, Y., AlRashed, S., Alsubaie, S., Almushayqih, Y., Mirza, F., Alotaibi, N., Al-Twairesh, N., Alowisheq, A., Bari, M. S., and Khan, H. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13787–13805, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.744. URL https://aclanthology.org/2024.acl-long.744/.

Amirizaniani, M., Lavergne, A., Snell Okada, E., Chadha, A., Roosta, T., and Shah, C. Developing a framework for auditing large language models using human-in-the-loop. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2025, pp. 64–74, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400722189. doi: 10.1145/3767695.3769514. URL https://doi.org/10.1145/3767695.3769514.

Ashktorab, Z., Geyer, W., Desmond, M., Daly, E. M., Cooper, M. S., Pan, Q., Miehling, E., Pedapati, T., and Do, H. J. Evalassist: A human-centered tool for llm-as-a-judge. *Workshop on Human-centered Evaluation and Auditing of Language Models*, 2025.

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

Bandy, J. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1):74:1–74:34, April 2021. doi: 10.1145/3449148. URL https://dl.acm.org/doi/10.1145/3449148.

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., Ghaleb, E., Giulianelli, M., Hanna, M., Koller, A., Martins, A., Mondorf, P., Neplenbroek, V., Pezzelle, S., Plank, B., Schlangen, D., Suglia, A., Surikuchi, A. K., Takmaz, E., and Testoni, A. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 238–255, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-252-7. doi: 10.18653/v1/2025.acl-short.20. URL https://aclanthology.org/2025.acl-short.20/.

Bengio, Y., McDermott, M., Cohen, M. K., Malkin, N., Fornasiere, D., Greiner, P., and Kaddar, Y. Can a bayesian oracle prevent harm from an agent? *SuperIntelligence-Robotics-Safety & Alignment*, 2(1), 2025.

Birhane, A., Steed, R., Ojewale, V., Vecchione, B., and Raji, I. D. Ai auditing: The broken bus on the road to ai accountability. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 612–643, 2024. doi: 10.1109/SaTML59370.2024.00037.

Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., and Wallach, H. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Inter-*

national Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 1004–1015, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.81. URL https://aclanthology.org/2021.acl-long.81/.

Brown, S., Davidovic, J., and Hasan, A. The algorithm audit: Scoring the algorithms that score us. *Big Data & Society*, 8(1):2053951720983865, 2021. doi: 10.1177/2053951720983865. URL https://doi.org/10.1177/2053951720983865.

Buolamwini, J. and Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pp. 77–91. PMLR, January 2018. URL https://proceedings.mlr.press/v81/buolamwini18a.html.

Cabrera, A. A., Druck, A. J., Hong, J. I., and Perer, A. Discovering and Validating AI Errors With Crowdsourced Failure Reports. *Proc. ACM Hum.-Comput. Interact.*, 5 (CSCW2):425:1–425:22, October 2021. doi: 10.1145/3479569. URL https://dl.acm.org/doi/10.1145/3479569.

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., Sharkey, L., Krishna, S., Von Hagen, M., Alberti, S., Chan, A., Sun, Q., Gerovitch, M., Bau, D., Tegmark, M., Krueger, D., and Hadfield-Menell, D. Black-Box Access is Insufficient for Rigorous AI Audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2254–2272, Rio de Janeiro Brazil, June 2024. ACM. ISBN 979-8-4007-0450-5. doi: 10.1145/3630106.3659037. URL https://dl.acm.org/doi/10.1145/3630106.3659037.

Cen, S. H. and Alur, R. From Transparency to Accountability and Back: A Discussion of Access and Evidence in AI Auditing. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '24, pp. 1–14, New York, NY, USA, October 2024. Association for Computing Machinery. ISBN 979-8-4007-1222-7. doi: 10.1145/3689904.3694711. URL https://doi.org/10.1145/3689904.3694711.

Chataigner, C., Ma, R., Ganesh, P., Chen, Y., Taïk, A., Creager, E., and Farnadi, G. Say It Another Way: Auditing LLMs with a User-Grounded Automated Paraphrasing Framework, October 2025. URL http://arxiv.org/abs/2505.03563. arXiv:2505.03563 [cs].

Chehbouni, K., Roshan, M., Ma, E., Wei, F., Taik, A., Cheung, J., and Farnadi, G. From representational harms to quality-of-service harms: A case study on llama 2 safety safeguards. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 15694–15710, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.927. URL https://aclanthology.org/2024.findings-acl.927.

Chehbouni, K., Haddou, M., Cheung, J. C., and Farnadi, G. Neither Valid nor Reliable? Investigating the Use of LLMs as Judges. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*, October 2025. URL https://openreview.net/forum?id=yqKfMr0yvY.

Chen, L., Zhang, Z., Tan, H., Dai, Q., Yang, H., Dong, Z., and Chen, X. Distributional LLM-as-a-judge. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=0SRGbRbngJ.

Cheong, I., Xia, K., Feng, K. J. K., Chen, Q. Z., and Zhang, A. X. (A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 2454–2469, New York, NY, USA, June 2024. Association for Computing Machinery. ISBN 979-8-4007-0450-5. doi: 10.1145/3630106.3659048. URL https://dl.acm.org/doi/10.1145/3630106.3659048.

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., and Smith, N. A. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7282–7296, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.565. URL https://aclanthology.org/2021.acl-long.565/.

Cohen-Inger, N., Elisha, Y., Shapira, B., Rokach, L., and Cohen, S. Forget what you know about llms evaluations – llms are like a chameleon, 2025. URL https://arxiv.org/abs/2502.07445.

Costanza-Chock, S., Raji, I. D., and Buolamwini, J. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 1571–1583, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. doi:

10.1145/3531146.3533213. URL https://dl.acm.org/doi/10.1145/3531146.3533213.

Crockett, M. J. and Messeri, L. AI Surrogates and illusions of generalizability in cognitive science. *Trends in Cognitive Sciences*, 0(0), October 2025. ISSN 1364-6613, 1879-307X. doi: 10.1016/j.tics.2025.09.012. URL https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(25)00251-7. Publisher: Elsevier.

Dahl, M., Magesh, V., Suzgun, M., and Ho, D. E. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis*, 16(1):64–93, January 2024. ISSN 2161-7201. doi: 10.1093/jla/laae003. URL https://doi.org/10.1093/jla/laae003.

Deng, W. H., Claire, W., Han, H. Z., Hong, J. I., Holstein, K., and Eslami, M. WeAudit: Scaffolding User Auditors and AI Practitioners in Auditing Generative AI. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–35, October 2025. ISSN 2573-0142. doi: 10.1145/3757702. URL https://dl.acm.org/doi/10.1145/3757702.

DeVos, A., Dhabalia, A., Shen, H., Holstein, K., and Eslami, M. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pp. 1–19, New York, NY, USA, April 2022. Association for Computing Machinery. ISBN 978-1-4503-9157-3. doi: 10.1145/3491102.3517441. URL https://dl.acm.org/doi/10.1145/3491102.3517441.

Esmaeilzadeh, P. Challenges and strategies for wide-scale artificial intelligence (ai) deployment in healthcare practices: A perspective for healthcare organizations. *Artificial Intelligence in Medicine*, 151:102861, 2024.

Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., and Radev, D. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021. doi: 10.1162/tacl_a_00373. URL https://aclanthology.org/2021.tacl-1.24/.

Fan, S., Barlas, P., Christoforou, E., Otterbacher, J., Sadiq, S., and Demartini, G. Socio-Economic Diversity in Human Annotations. In *Proceedings of the 14th ACM Web Science Conference 2022*, WebSci '22, pp. 98–109, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9191-7. doi: 10.1145/3501247.3531588. URL https://dl.acm.org/doi/10.1145/3501247.3531588.

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Fei, Z., Shen, X., Zhu, D., Zhou, F., Han, Z., Huang, A., Zhang, S., Chen, K., Yin, Z., Shen, Z., Ge, J., and Ng, V. LawBench: Benchmarking legal knowledge of large language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7933–7962, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.452. URL https://aclanthology.org/2024.emnlp-main.452/.

Floridi, L., Holweg, M., Taddeo, M., Amaya, J., Mökander, J., and Wen, Y. capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act, March 2022. URL https://papers.ssrn.com/abstract=4064091.

Gadiraju, V., Kane, S., Dev, S., Taylor, A., Wang, D., Denton, R., and Brewer, R. "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pp. 205–216, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 979-8-4007-0192-4. doi: 10.1145/3593013.3593989. URL https://dl.acm.org/doi/10.1145/3593013.3593989.

Gorle, A. R., Yadav, A. K. S., and Weissman, T. Quantifying Information Gain and Redundancy in Multi-Turn LLM Conversations. In *First Workshop on Multi-Turn Interactions in Large Language Models*, November 2025. URL https://openreview.net/forum?id=5gpABTkcUJ.

Guerdan, L., Barocas, S., Holstein, K., Wallach, H., Wu, S., and Chouldechova, A. Validating LLM-as-a-judge systems under rating indeterminacy. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=ZwDMrArTBg.

Guha, N., Nyarko, J., Ho, D., Ré, C., Chilton, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D., Zambrano, D., et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279, 2023.

Guha, N., Lawrence, C. M., Gailmard, L. A., Rodolfa, K. T., Surani, F., Bommasani, R., Raji, I. D., Cuéllar, M.-F., Honigsberg, C., Liang, P., et al. Ai regulation has its own alignment problem: The technical and institutional feasibility of disclosure, registration, licensing, and auditing. *Geo. Wash. L. Rev.*, 92:1473, 2024.

Gupta, K. *Contemporary Auditing*. McGraw-Hill Education (India) Pvt Limited, 2004. ISBN 9780070585843. URL https://books.google.ca/books?id=neDFWDyUWuQC.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Hofmann, V., Heineman, D., Magnusson, I., Lo, K., Dodge, J., Sap, M., Koh, P. W., Wang, C., Hajishirzi, H., and Smith, N. A. Fluid language model benchmarking. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=mxcCg9YRqj.

Howcroft, D. M., Belz, A., Clinciu, M.-A., Gkatzia, D., Hasan, S. A., Mahamood, S., Mille, S., van Miltenburg, E., Santhanam, S., and Rieser, V. Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions. In Davis, B., Graham, Y., Kelleher, J., and Sripada, Y. (eds.), *Proceedings of the 13th International Conference on Natural Language Generation*, pp. 169–182, Dublin, Ireland, December 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.inlg-1.23. URL https://aclanthology.org/2020.inlg-1.23/.

Hu, X., Gao, M., Hu, S., Zhang, Y., Chen, Y., Xu, T., and Wan, X. Are LLM-based evaluators confusing NLG quality criteria? In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9530–9570, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.516. URL https://aclanthology.org/2024.acl-long.516/.

Hupont, I., Micheli, M., Delipetrev, B., Gómez, E., and Garrido, J. S. Documenting high-risk ai: a european regulatory perspective. *Computer*, 56(5):18–27, 2023.

IEEE. IEEE Standard for Software Reviews and Audits. *IEEE Std 1028-2008*, pp. 1–53, August 2008. doi: 10.1109/IEEESTD.2008.4601584. URL https://ieeexplore.ieee.org/document/4601584.

Jain, N., Han, K., Gu, A., Li, W.-D., Yan, F., Zhang, T., Wang, S., Solar-Lezama, A., Sen, K., and Stoica, I. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=chfJJYC3iL.

Jiang, L., Chai, Y., Li, M., Liu, M., Fok, R., Dziri, N., Tsvetkov, Y., Sap, M., and Choi, Y. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=saDOrrnNTz.

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

Kapania, S., Taylor, A. S., and Wang, D. A hunt for the Snark: Annotator Diversity in Data Practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pp. 1–15, New York, NY, USA, April 2023. Association for Computing Machinery. ISBN 978-1-4503-9421-5. doi: 10.1145/3544548.3580645. URL https://dl.acm.org/doi/10.1145/3544548.3580645.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2564–2572. PMLR, July 2018. URL https://proceedings.mlr.press/v80/kearns18a.html. ISSN: 2640-3498.

Kim, E., Suk, J., Kim, S., Muennighoff, N., Kim, D., and Oh, A. LLM-as-an-Interviewer: Beyond Static Testing Through Dynamic LLM Evaluation. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26456–26493, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1357. URL https://aclanthology.org/2025.findings-acl.1357/.

Kuhn, L., Gal, Y., and Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Laban, P., Hayashi, H., Zhou, Y., and Neville, J. Llms get lost in multi-turn conversation, 2025. URL https://arxiv.org/abs/2505.06120.

Lee, J., Alvero, A., Joachims, T., and Kizilcec, R. F. Poor alignment and steerability of large language models: Evidence using 30,000 college admissions essays. In *Workshop on Socially Responsible Language Modelling Research*, 2025a. URL https://openreview.net/forum?id=LTRDBQWVe4.

Lee, J., Kim, S., Han, J., Lee, J.-M., Kim, K., Oh, A., and Choi, E. Trans-env: A framework for evaluating the linguistic robustness of LLMs against english varieties. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025b. URL https://openreview.net/forum?id=YIpvHrQAks.

Li, S. S., Balachandran, V., Feng, S., Ilgen, J. S., Pierson, E., Koh, P. W., and Tsvetkov, Y. MediQ: Question-Asking LLMs and a Benchmark for Reliable Interactive Clinical Reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888, December 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/32b80425554e081204e5988ab1c97e9a-Abstract-Conference.html.

Li, X., Lan, Y., and Yang, C. TreeEval: Benchmark-Free Evaluation of Large Language Models through Tree Planning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24485–24493, April 2025a. ISSN 2374-3468. doi: 10.1609/aaai.v39i23.34627. URL https://ojs.aaai.org/index.php/AAAI/article/view/34627.

Li, Y. and Goel, S. Making It Possible for the Auditing of AI: A Systematic Review of AI Audits and AI Auditability. *Information Systems Frontiers*, 27(3):1121–1151, June 2025. ISSN 1572-9419. doi: 10.1007/s10796-024-10508-8. URL https://doi.org/10.1007/s10796-024-10508-8.

Li, Y., Xiong, M., Wu, J., and Hooi, B. Conftuner: Training large language models to express their confidence verbally. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=VZQ04Ojhu5.

Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. G-eval: NLG evaluation using gpt-4 with better human alignment. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Liu, Y., Moosavi, N., and Lin, C. LLMs as narcissistic evaluators: When ego inflates evaluation scores. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12688–12701, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.753. URL https://aclanthology.org/2024.findings-acl.753/.

Liu, Y. L., Blodgett, S. L., Cheung, J., Liao, Q. V., Olteanu, A., and Xiao, Z. ECBD: Evidence-Centered Benchmark Design for NLP. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16349–16365, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.861. URL https://aclanthology.org/2024.acl-long.861/.

Lunardi, R., Della Mea, V., Mizzaro, S., and Roitero, K. On robustness and reliability of benchmark-based evaluation of llms. In *ECAI 2025*, pp. 4603–4610. IOS Press, 2025.

Megan Ma. Legal Training Should Embrace Generative AI Large Language Models, December 2023. URL https://news.bloomberglaw.com/us-law-week/legal-training-should-embrace-generative-ai-large-language-models.

Metaxa, D., Park, J. S., Robertson, R. E., Karahalios, K., Wilson, C., Hancock, J., and Sandvig, C. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends in Human-Computer Interaction*, 14(4):272–344, November 2021. ISSN 1551-3955. doi: 10.1561/1100000083. URL https://doi.org/10.1561/1100000083.

Minkkinen, M., Laine, J., and Mäntymäki, M. Continuous Auditing of Artificial Intelligence: a Conceptualization and Assessment of Tools and Frameworks. *Digital Society*, 1(3):21, October 2022. ISSN 2731-4669. doi: 10.1007/s44206-022-00022-2. URL https://doi.org/10.1007/s44206-022-00022-2.

Miranda, L. J. V., Wang, Y., Elazar, Y., Kumar, S., Pyatkin, V., Brahman, F., Smith, N. A., Hajishirzi, H., and

Dasigi, P. Hybrid preferences: Learning to route instances for human vs. AI feedback. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7162–7200, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.355. URL https://aclanthology.org/2025.acl-long.355/.

Mökander, J., Morley, J., Taddeo, M., and Floridi, L. Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. *Science and Engineering Ethics*, 27(4):44, July 2021. ISSN 1471-5546. doi: 10.1007/s11948-021-00319-4. URL https://doi.org/10.1007/s11948-021-00319-4.

Mökander, J., Schuett, J., Kirk, H. R., and Floridi, L. Auditing large language models: A three-layered approach. *AI and Ethics*, 4(4):1085–1115, November 2024. ISSN 2730-5961. doi: 10.1007/s43681-023-00289-2. URL https://doi.org/10.1007/s43681-023-00289-2.

Neumann, A. and Singh, J. Caught in the Cascade: Why LLM Auditing is Missing the Middle. *Workshop on Human-centered Evaluation and Auditing of Language Models*, 2025.

Palta, R., Angwin, J., and Nelson, A. How We Tested Leading AI Models Performance on Election Queries, February 2024. URL https://www.proofnews.org/how-we-tested-leading-ai-models-performance-on-election-queries/.

Pan, J., Shar, R., Pfau, J., Talwalkar, A., He, H., and Chen, V. When Benchmarks Talk: Re-Evaluating Code LLMs with Interactive Feedback. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24672–24700, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1267. URL https://aclanthology.org/2025.findings-acl.1267/.

Pan, Q., Ashktorab, Z., Desmond, M., Santillán Cooper, M., Johnson, J., Nair, R., Daly, E., and Geyer, W. Human-centered design recommendations for LLM-as-a-judge. In Soni, N., Flek, L., Sharma, A., Yang, D., Hooker, S., and Schwartz, H. A. (eds.), *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pp. 16–29, TBD, August 2024. ACL. doi: 10.18653/v1/2024.hucllm-1.2. URL https://aclanthology.org/2024.hucllm-1.2/.

Parrish, A., Prabhakaran, V., Aroyo, L., Díaz, M., Homan, C. M., Serapio-García, G., Taylor, A. S., and Wang, D. Diversity-Aware Annotation for Conversational AI Safety. In Dinkar, T., Attanasio, G., Cercas Curry, A., Konstas, I., Hovy, D., and Rieser, V. (eds.), *Proceedings of Safety4ConvAI: The Third Workshop on Safety for Conversational AI @ LREC-COLING 2024*, pp. 8–15, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.safety4convai-1.2/.

Peeperkorn, M., Kouwenhoven, T., Brown, D., and Jordanous, A. Is temperature the creativity parameter of large language models? In *ICCC*, pp. 226–235, 2024. URL https://computationalcreativity.net/iccc24/papers/ICCC24_paper_70.pdf.

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Zhang, C. B. C., Shaaban, M., Ling, J., Shi, S., and others. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Polo, F. M., Wang, X., Yurochkin, M., Xu, G., Banerjee, M., and Sun, Y. Bridging human and LLM judgments: Understanding and narrowing the gap. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=bEP87LNTfX.

Prabhudesai, S., Kasi, A. P., Mansingh, A., Das Antar, A., Shen, H., and Banovic, N. "Here the GPT made a choice, and every choice can be biased": How Students Critically Engage with LLMs through End-User Auditing Activity. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pp. 1–23, New York, NY, USA, April 2025. Association for Computing Machinery. ISBN 979-8-4007-1394-1. doi: 10.1145/3706598.3713714. URL https://dl.acm.org/doi/10.1145/3706598.3713714.

Raji, I. D. and Buolamwini, J. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society*, Aies '19, pp. 429–435, Honolulu, HI, USA, 2019. Association for Computing Machinery. ISBN 978-1-4503-6324-2. doi: 10.1145/3306618.3314244. URL https://doi.org/10.1145/3306618.3314244. Number of pages: 7 tex.address: New York, NY, USA.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 33–44, New York, NY, USA, January 2020. Association for Computing Machinery. ISBN 978-1-4503-6936-7. doi:

10.1145/3351095.3372873. URL https://dl.acm.org/doi/10.1145/3351095.3372873.

Rastogi, C., Tulio Ribeiro, M., King, N., Nori, H., and Amershi, S. Supporting Human-AI Collaboration in Auditing LLMs with LLMs. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 913–926, Montréal QC Canada, August 2023. ACM. ISBN 979-8-4007-0231-0. doi: 10.1145/3600211.3604712. URL https://dl.acm.org/doi/10.1145/3600211.3604712.

Rawte, V., Sheth, A., and Das, A. A survey of hallucination in large foundation models, 2023. URL https://arxiv.org/abs/2309.05922.

Ribeiro, M. T. and Lundberg, S. Adaptive Testing and Debugging of NLP Models. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3253–3267, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.230. URL https://aclanthology.org/2022.acl-long.230/.

Salaudeen, O. E., Reuel, A., Ahmed, A. M., Bedi, S., Robertson, Z., Sundar, S., Domingue, B. W., Wang, A., and Koyejo, S. Measurement to Meaning: A Validity-Centered Framework for AI Evaluation. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, September 2025. URL https://openreview.net/forum?id=2Bw6uC49QF&referrer=%5Bthe%20profile%20of%20Sanmi%20Koyejo%5D(%2Fprofile%3Fid%3D~Sanmi_Koyejo1).

Sandvig, C., Hamilton, K., Karahalios, K., and Langbort, C. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 22(2014):4349–4357, 2014.

Schuff, H., Vanderlyn, L., Adel, H., and Vu, N. T. How to do human evaluation: A brief introduction to user studies in NLP. *Natural Language Engineering*, 29(5):1199–1222, 2023. doi: 10.1017/S1351324922000535.

Selbst, A. D. An institutional view of algorithmic impact assessments. *Harv. JL & Tech.*, 35:117, 2021.

Shankar, S., Zamfirescu-Pereira, J., Hartmann, B., Parameswaran, A., and Arawjo, I. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706288. doi: 10.1145/3654777.3676450. URL https://doi.org/10.1145/3654777.3676450.

Shen, H., Knearem, T., Ghosh, R., Alkiek, K., Krishna, K., Liu, Y., Petridis, S., Peng, Y.-H., Qiwei, L., Si, C., Xie, Y., Bigham, J. P., Bentley, F., Chai, J., Lipton, Z. C., Mei, Q., Terry, M., Yang, D., Morris, M. R., Resnick, P., and Jurgens, D. Position: Towards bidirectional human-AI alignment. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*, 2025. URL https://openreview.net/forum?id=PgA9rZoMY8.

Si, C., Goyal, N., Wu, T., Zhao, C., Feng, S., Daumé Iii, H., and Boyd-Graber, J. Large language models help humans verify truthfulness – except when they are convincingly wrong. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1459–1474, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.81. URL https://aclanthology.org/2024.naacl-long.81/.

Siska, C., Marazopoulou, K., Ailem, M., and Bono, J. Examining the robustness of LLM evaluation to the distributional assumptions of benchmarks. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10406–10421, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.560. URL https://aclanthology.org/2024.acl-long.560.

Sourati, Z., Karimi-Malekabadi, F., Ozcan, M., McDaniel, C., Ziabari, A., Trager, J., Tak, A., Chen, M., Morstatter, F., and Dehghani, M. The shrinking landscape of linguistic diversity in the age of large language models, 2025. URL https://arxiv.org/abs/2502.11266.

Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P., and Langer, M. On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 2495–2507, New York, NY, USA, June 2024. Association for Computing Machinery. ISBN 979-8-4007-0450-5. doi: 10.1145/3630106.3659051. URL https://dl.acm.org/doi/10.1145/3630106.3659051.

Suresh, H. and Guttag, J. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms,* *and Optimization*, EAAMO '21, pp. 1–9. ACM, October 2021. doi: 10.1145/3465416.3483305. URL http://dx.doi.org/10.1145/3465416.3483305.

Troshin, S., Mohammed, W., Meng, Y., Monz, C., Fokkens, A., and Niculae, V. Control the temperature: Selective sampling for diverse and high-quality LLM outputs. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=IyOC5GCzv4.

Truong, S. T., Tu, Y., Liang, P., Li, B., and Koyejo, S. Reliable and Efficient Amortized Model-based Evaluation. In *Forty-Second International Conference on Machine Learning*, June 2025. URL https://openreview.net/forum?id=HDbWrsgkB9&referrer=%5Bthe%20profile%20of%20Percy%20Liang%5D(%2Fprofile%3Fid%3D~Percy_Liang1).

Wallach, H., Desai, M., Cooper, A. F., Wang, A., Atalla, C., Barocas, S., Blodgett, S. L., Chouldechova, A., Corvi, E., Dow, P. A., Garcia-Gathright, J., Olteanu, A., Pangakis, N. J., Reed, S., Sheng, E., Vann, D., Vaughan, J. W., Vogel, M., Washington, H., and Jacobs, A. Z. Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge. In *Forty-Second International Conference on Machine Learning Position Paper Track*, June 2025. URL https://openreview.net/forum?id=1ZC4RNjqzU.

Wang, A., Morgenstern, J., and Dickerson, J. P. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3):400–411, March 2025a. ISSN 2522-5839. doi: 10.1038/s42256-025-00986-z. URL https://doi.org/10.1038/s42256-025-00986-z.

Wang, C., Chen, Z., Li, T., Zhang, Y., and Liu, Y. Towards Trustworthy LLMs for Code: A Data-Centric Synergistic Auditing Framework. In *2025 IEEE/ACM 47th International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)*, pp. 56–60, Ottawa, ON, Canada, April 2025b. IEEE. ISBN 979-8-3315-3711-1. doi: 10.1109/ICSE-NIER66352.2025.00017. URL https://ieeexplore.ieee.org/document/11023943/.

Wang, R., Zhang, Q., Robinson, C., Loeb, S., and Demszky, D. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2174–2199, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.120. URL https://aclanthology.org/2024.naacl-long.120/.

Wang, X., Wang, Z., Liu, J., Chen, Y., Yuan, L., Peng, H., and Ji, H. MINT: Evaluating LLMs in multi-turn interaction with tools and language feedback. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=jp3gWrMuIZ.

Wei, F., Chen, X., and Luo, L. Rethinking generative large language model evaluation for semantic comprehension. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52525–52558. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wei24c.html.

Weidinger, L., Raji, I. D., Wallach, H., Mitchell, M., Wang, A., Salaudeen, O., Bommasani, R., Ganguli, D., Koyejo, S., and Isaac, W. Toward an evaluation science for generative ai systems, 2025. URL https://arxiv.org/abs/2503.05336.

Weiser, B. Here's What Happens When Your Lawyer Uses ChatGPT. *The New York Times*, May 2023. ISSN 0362-4331. URL https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html.

Wright, D., Masud, S., Moore, J., Yadav, S., Antoniak, M., Park, C. Y., and Augenstein, I. Epistemic Diversity and Knowledge Collapse in Large Language Models, October 2025. URL http://arxiv.org/abs/2510.04226. arXiv:2510.04226 [cs].

Wu, F., Black, E., and Chandrasekaran, V. Generative monoculture in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=yZ7sn9pyqb.

Yadkori, Y. A., Kuzborskij, I., György, A., and Szepesvári, C. To believe or not to believe your llm. *arXiv preprint arXiv:2406.02543*, 2024.

Yu, Z., Gao, C., Yao, W., Wang, Y., Ye, W., Wang, J., Xie, X., Zhang, Y., and Zhang, S. KIEval: A Knowledge-grounded Interactive Evaluation Framework for Large Language Models. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5967–5985, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

doi: 10.18653/v1/2024.acl-long.325. URL https://aclanthology.org/2024.acl-long.325/.

Zhang, J., Yu, S., Chong, D., Sicilia, A., Tomz, M. R., Manning, C. D., and Shi, W. Verbalized Sampling: How to Mitigate Mode Collapse and Unlock LLM Diversity, October 2025a. URL http://arxiv.org/abs/2510.01171. arXiv:2510.01171 [cs].

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Zhang, Y., Diddee, H., Holm, S., Liu, H., Liu, X., Samuel, V., Wang, B., and Ippolito, D. Noveltybench: Evaluating creativity and diversity in language models. In *Second Conference on Language Modeling*, 2025b. URL https://openreview.net/forum?id=XZm1ekzERf.

Zhao, W., Ren, X., Hessel, J., Cardie, C., Choi, Y., and Deng, Y. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

Zhou, K., Blodgett, S. L., Trischler, A., Daumé III, H., Suleman, K., and Olteanu, A. Deconstructing NLG Evaluation: Evaluation Practices, Assumptions, and Their Implications. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V. (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 314–324, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.24. URL https://aclanthology.org/2022.naacl-main.24/.

Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W. X., Chen, X., Lin, Y., Wen, J.-R., and Han, J. Don't make your llm an evaluation benchmark cheater, 2023. URL https://arxiv.org/abs/2311.01964.

Zhu, K., Chen, J., Wang, J., Gong, N. Z., Yang, D., and Xie, X. DyVal: Dynamic Evaluation of Large Language Models for Reasoning Tasks. In *The Twelfth International Conference on Learning Representations*, October 2023. URL https://openreview.net/forum?id=gjfOL9z5Xr.

A. AI Auditing dimensions

Table 1 summarizes key dimensions along which AI audits can vary, highlighting how differences in objectives, auditor position, system access, relationships with developers, and timing, shape what evidence can be collected and what conclusions can be drawn.

While many tools discussed in this paper apply across all these axes, our primary focus is on *compliance-based audits*, which assess adherence to predefined standards rather than *risk-based audits* aimed at informing model development. We further restrict our analysis to *black-box auditing*.

Table 1. Key dimensions along which AI audits differ (Mökander et al., 2024; Sandvig et al., 2014; Raji et al., 2020; Casper et al., 2024).

Dimension	**Type**	**Definition**
Audit Objective	Compliance-based	Assesses adherence to predefined rules, standards, or legal requirements (e.g., regulations, internal policies, documentation obligations).
	Risk-based	Asks open-ended questions about how a system works to identify and control risks.
Auditor Position	Internal	Conducted by the system developer or deploying organization, often integrated into internal governance, risk management, or quality assurance processes. Allows deeper system access but may raise independence concerns.
	External	Performed by independent third parties (e.g., regulators, civil society, auditors, researchers). Typically operates under constrained access to systems and data.
System Access	Black-box	Allows auditors to design inputs for a system, query it, and analyze the resulting outputs.
	White-box	Allows auditors full access to the system. This includes access to weights, activations, gradients, and the ability to fine-tune the model.
	Outside-the-box	Grants auditors access to additional information about the system's development and deployment. There are many types, which can include methodological details, source code, documentation, hyperparameters, training data, deployment details, and findings from internal evaluations.
Auditor–Developer Relationship	Adversarial	Auditors operate without cooperation from the system owner, often using black-box testing, probing, or reverse engineering to uncover risks or harms.
	Collaborative	Auditors and system developers cooperate, sharing internal documentation, models, or data to jointly identify and mitigate risks.
Audit Timing	Ex-ante	Conducted before system deployment to anticipate and mitigate potential risks.
	Ex-post	Conducted after deployment, drawing on real-world performance, observed harms, user feedback, and incident reports.