Make Mechanistic Interpretability Auditable: A Call to Develop Guidelines via Continuous Collaborative Reviewing

Michael Lan Martian
Narmeen Fatimah Oozeer Martian
Chaithanya Bandi Martian
Philip Quirke Martian
Austin Meek University of Delaware, MATS
Fazl Barez University of Oxford, Martian
Amirali Abdullah ThoughtWorks, Martian

Abstract

While mechanistic interpretability (MI) has produced important insights into neural network internals, the field has yet to establish a standardized system to audit experiments. As such, many of its findings remain underutilized in safety-critical applications such as medical AI and autonomous systems, as stakeholders cannot certify their validity. Recent work demonstrates this concretely: two papers found conflicting conclusions for the same behavior, and a third study revealed that both were partially correct but incomparable due to methodological inconsistencies. Without standardized auditing, such ambiguities hinder adoption in high-stakes contexts requiring strong correctness guarantees. We call for the MI community to work towards developing a novel reviewing system that complements peer review via: (1) Continuous reviewing supported by a *Collaborative Reviewing Platform* where meta-science results and discussions (such as critiques, negative results, post-hoc extensions, reproductions, replications, and partial results) that fit outside of papers are organized and discussed, allowing for comments and revisions to be made at any time (2) Generalizing good practices found on this platform into expert-verified guidelines and protocols to improve auditing efficiency, and (3) Source-based auditing systems that track arguments which claims depend on. This position paper encourages constructive debate over the necessity, design and implementation of such a framework, providing early concrete examples to help catalyze these dialogues. Overall, we propose that auditing MI itself is essential for its application in AI safety, industry, and governance.

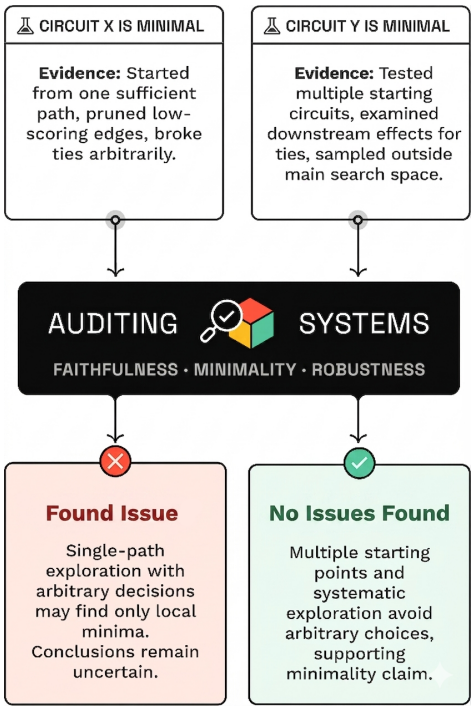

Figure 1: A high-level example of auditing two different hypothesized claims based on community-refined guidelines.

1 Introduction

Two mechanistic interpretability studies proposed competing explanations for the same mechanism in a neural network (Chughtai et al., 2023; Stander et al., 2024). Each was peer reviewed, and both appeared credible. Yet, they reached contradictory conclusions about the internal mechanisms responsible for the model's behavior. Only when a third paper reconciled them under a unifying framework was it revealed that both were partially correct, yet incomparable due to their experimental methodology (Wu et al., 2025b). For stakeholders considering deployment in safety-critical applications such as medical diagnostics, autonomous systems, and financial regulation, such ambiguity is unacceptable. Without good auditing protocols, how can decision-makers determine which claims to trust?

Mechanistic interpretability (MI), the study of discovering algorithmic explanations for the internal workings of neural networks (Bereska and Gavves, 2024), has advanced rapidly in recent

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3516–3540

years, producing important insights for practical applications such as model steering (Anthropic, 2025), hallucination detection (Obeso et al., 2025), and AI auditing (Marks et al., 2025). These advances have generated significant interest across academia and industry (Amodei, 2025; Balsam et al., 2024). However, the field has grown without establishing widely enforced standards for auditing experiments. This risks serious issues: leading MI experts (Sharkey et al., 2025) warn that without extensive hypothesis validation, researchers are in danger of conducting studies with significant room for subjective interpretation, lacking objective frameworks to adjudicate competing explanations. This increases the chances of overconfident conclusions and "interpretability illusions": claims which appear correct, yet fail essential sanity checks (Adebayo et al., 2018; Friedman et al., 2024).

These systemic issues have dire consequences for the field, making its findings less likely to be adopted for real-world applications that require strong safety guarantees (Wu et al., 2025a; Golgoon et al., 2024). While studies from leading experts benefit from rigorous internal verification (Nanda et al., 2023; Tigges et al., 2024), the broader research community lacks this assurance. Without certification procedures, valuable findings remain underutilized, and if these issues are not corrected, the field is at risk of being overwhelmed by studies of uncertain quality.

The MI community has recognized this challenge. Bereska and Gavves (2024) emphasizes that objective evaluation requires well-defined metrics, standardized benchmarks, and algorithmic testbeds. Casper (2023) argues that MI must adopt engineering rigor for grounding claims. Méloux et al. (2025a) argues that AI interpretability methods can produce plausible-looking explanations even for spurious or weakly supported patterns, and therefore reframes interpretability as a statistical inference problem that should quantify uncertainty and test explanations against explicit alternatives. Recent work argues that MI explanations may be fundamentally non-identifiable, with multiple distinct circuits or abstractions equally matching the same behavior (Méloux et al., 2025b). Recent theory also suggests that causal abstraction, without additional restrictions on the complexity or form of alignment maps, can become too permissive to be informative, implying that MI audits must evaluate not just fit but the assumptions that make an explanation meaningful (Sutter et al., 2025). Second Look Research aims to produced open-source replications of MI and AI safety research in a centralized manner, addressing reproducibility issues in research (Semmelrock et al., 2025). Yet despite these urgent calls, a widely accepted and consistently enforced approach has yet to be established.

Due to rapidly evolving research, such that previously declared validity can vastly change in short time periods, we advocate for the research community to pursue the development of **continuous reviewing** that complements peer review. We define continuous reviewing as a collaborative approach that gradually refines pieces of research using **meta-analysis** results and discussions that fit outside of a paper. Meta-analysis includes comments, criticisms, negative results, reproductions, replications, non-novel extensions, minor yet useful results which are not enough for a paper, and partial results. We define these results that complement existing papers but do not exist in paper-form as **meta-results**. We propose that this development will benefit via improved organization of meta-results and discussions using a **community-driven platform**. We posit continuous reviewing as an experimental approach following previous works such as OpenReview (Soergel et al., 2013). Useful auditing patterns found on this platform from refining meta-results can eventually transform into standardized empirical guidelines.

This paper's aim is not to give specific guidelines, but to advocate for the development of an auditing system for MI, arguing this can be done by first improving meta-analysis organization. We propose a concrete implementation roadmap via a community-driven reviewing system built with three important components:

1. **Continuous Reviewing using Meta-Analysis** supported by a *Collaborative Reviewing Platform* where meta-results and discussions (such as negative results, reproductions, and partial results) that fit outside papers are neatly organized and built upon by researchers, allowing for comments and revisions to be made at any time.

2. **Community-Refined Guidelines and Protocols** generalized from common, useful and expert-supported points developed on this platform that assist in both community reviewing and professional auditing.

3. **Source-Based Auditing Systems** that weigh claims by tracing and tracking the assumptions, evidence, specific experiments, and other claims they depend on.

Crucially, we emphasize that any developed standards should not be seen as the most definitive test of a study's quality; rather, we advocate that they provide one rigorous dimension, out of many, to include in nuanced discussions during study evaluations. This approach helps avoid issues such as a false sense of rigor when a study passes a checklist. Figure 1 gives an auditing process example.

This position paper seeks to spark constructive discussion about the necessity, design, and implementation of an auditing framework for MI, introducing concrete examples to help catalyze dialogues, and arguing for continuous community reviewing. By establishing this framework early, we ensure that MI matures into a reliable approach for AI safety, rather than remaining an informal practice. Our main contributions include:

- Arguments establishing the need for MI auditing, using examples to motivate discussions
- A plan-to-action for implementing a community-based MI reviewing system, proposing an approach using continuous reviewing via meta-analysis
- A proposal for a collaborative platform where researchers propose community-refined guidelines which influence professional auditing guidelines in safety-critical applications

2 Technical Background and Potential MI Pitfalls

Due to the inherent difficulty of explaining black-box models, MI studies may fall into empirical pitfalls that span across approaches, which can be avoided by following guidelines as described in Table 1. However, each approach also has its own unique pitfalls, as shown in Table 2 . These complexities highlight the need to develop a wide range of guidelines. Given the broad scale of MI, for explanatory purposes, we explain technical reasons that lead to some pitfalls for one approach: *Activation Patching for Circuit Discovery* (Meng et al., 2022).

We define model M as a directed graph whose nodes correspond to internal components (e.g., attention heads), and whose edges represent information flow. A *circuit* $C \subset M$ is a subgraph hypothesized to implement a specific behavior, such as induction. We define a metric $F_D(M, C)$ to compare the behaviors of C to M on dataset D.

Activation patching replaces a clean prompt into a corrupted one to measure how much performance is restored. Let a_k^{clean} and a_k^{corr} be the activations at component k, and define a patched model $M^{(k)}$ by replacing only that component:

$$M^{(k)}_{\text{patched}}(x_{\text{corr}}) = M\big(x_{\text{corr}} \text{ with } a_k^{\text{clean}} \to a_k^{\text{corr}}\big).$$

A typical score then compares a metric F (such as a logit-difference) before and after patching:

$$\Delta_F(k) = F\big(M^{(k)}_{\text{patched}}(x_{\text{corr}})\big) - F\big(M(x_{\text{corr}})\big).$$

Large values of $\Delta_F(k)$ are usually interpreted as evidence that k contributes to the behavior associated with the clean prompt. This is applied over multiple components to localize a circuit.

Hypothesis-testing formulations note that such scores alone do not ensure that a proposed circuit is well-defined (Shi et al., 2025). Even if $\Delta_F(k)$ is large for all k in a candidate subgraph C, one must also check that each component is necessary. This is often expressed as a *minimality* condition:

$$\forall\, k \in C: \quad F_D\big(M,\, C \setminus \{k\}\big) > F_D(M, C),$$

Because small choices in metrics or circuit reduction can produce wildly different attributions that look equally plausible, practitioners can be mislead into overconfidence in their claims. Corruption schemes can push representations off-distribution, producing spurious $\Delta_F(k)$. Metric selection (e.g., logits vs. probabilities) can flip the ranking of important components. Therefore, given the sensitivities of experimental conduct to study quality, it is paramount to standardize nuanced guidelines, especially as methods become more complex.

Additionally, we present an example of using guidelines to audit an MI study in a safety-critical medical scenario in Appendix B. These examples illustrate that evaluating MI experiments often requires careful inspection of methodological details that are not always visible from headline results alone. While experienced practitioners may recognize such issues informally, there is currently no standardized process for recording, comparing, and auditing these practices across studies, which is essential for AI safety applications described in Appendix E.

Pitfall	Description	Auditing Guideline
Interpretability illusions (Friedman et al., 2024)	Plausible explanations that fail when evaluated on broader or counterfactual data.	Stress-test with counterfactual inputs and hold-out distributions
Cherry-picking (Casper, 2023)	Demonstrating results only on favorable tasks or hand-picked examples.	Evaluate on random or full data slices and report negative cases
Missing sanity checks (Adebayo et al., 2018)	Failing to test against random models, scrambled labels, or weight shuffling.	Run method on null baselines and verify degradation
No causal validation (Mueller et al., 2025a)	Relying solely on correlations or descriptive claims without interventions.	Use ablations, activation patching, or necessity/sufficiency tests

Table 1: Examples of potential pitfalls in MI experiments, and guidelines to audit whether experiments avoid them. Auditing checks if these guidelines are followed. Currently, these guidelines are high-level; we expect that as the field progresses and notices patterns in what works, it will develop more nuanced guidelines. More examples are given in Table 3 in Appendix C.

Approach	Potential Pitfalls	Auditing Guideline
Probing: Test whether a concept is encoded in activations	Correlation mistaken for causation (Belinkov, 2021)	Follow with causal interventions
Activation Patching: Test causality via interventions	Corrupted prompt choice changes outcomes (Zhang and Nanda, 2024)	Use multiple prompt distributions and track outcome variations
Sparse Decomposition: Decompose polysemantic activations (Elhage et al., 2022) into interpretable features	Claim a feature represents a concept, yet may fail to fire on some cases due to absorption (Chanin et al., 2025)	Validate coverage by checking recall, and use causal intervention to verify that failures are not due to feature absorption
Activation Steering: Manipulate activations to steer output	Steering may degrade under OOD distribution shifts (Tan et al., 2025)	Evaluate steering on diverse benchmarks

Table 2: Approach-Specific Auditing Guidelines. These are high-level guidelines that are well-known; however, more complex scenarios will call for more nuanced guidelines that the community should be aware of. These issues may be grouped together under a standardized taxonomy (e.g., some are types of interpretablity illusions).

3 Call for Community-Driven Empirical Guidelines

To ensure a clear understanding of MI study quality, it is essential to develop effective guidelines. However, given that MI is an emerging field, what constitutes as “good standards" is still being formulated, and lacks widespread agreement. But if no effort is being made to establish expert-recommended guidelines, the field risks becoming disorganized, confused, and overwhelmed by practices of uncertain quality, such that time and resources would be spent on studies which do not reach their full potential due to a lack of guidance.

Thus, to facilitate the development of effective auditing guidelines, we propose an organized effort to set up a *Collaborative Meta-Analysis Platform* consisting of: (1) Experiment repositories from which the community can organize, review, and cross-compare meta-results, and (2) Forums where users can propose and debate both claims and community-driven “living document" guidelines akin to Wikipedia[1], supporting positions with repository-hosted evidence and structured hypotheses chains.

This platform design assumes that certain guidelines should be backed not just by conceptual justifications, but by real-world empirical evidence.

[1]https://en.wikipedia.org/wiki/Living_document

We contend that such an open ecosystem will help ensure that standards are supported by transparent and evidence-based justifications, rather than being established by potentially corruptible authorities based on opaque reasons. By allowing guidelines to be open to revision, we avoid locking in assumptions too early in the field's development.

3.1 Motivation for Community-Driven Guideline Development

Current Landscape of MI. This paper is not a criticism of how the MI community currently operates; we believe that it is beneficial for a new field to start off as free-form and flexible. At present, it is still unclear what counts as "good experiment practices", and so innovative exploration should be the norm. However, after years of progress and many successful results, we believe that enough experience has been accumulated for the field to begin moving towards instituting experiment standards.

Individual MI researchers and organizations have produced a range of valuable educational resources, including well-written explanations of good experiment practices (Nanda, 2025). Publically-available training curriculum like ARENA have systematized empirical methods.[2] These resources have been very helpful for teaching practioners how to properly conduct MI research. However, they do not address the systematic auditing of experimental studies. Therefore, we propose that additional resources are needed to more explicitly lay out good experimental practices, both to guide researchers during a project, and to enable systematic post-hoc auditing.

The Open and Online Communities of MI. Many advances in MI research have been driven by its open and collaborative online communities (Saphra and Wiegreffe, 2024), such as forums[3] and Discord servers[4] where users can engage with MI experts. Additionally, MI research is highly accessible, as it can be done with fewer resources in comparison to other fields. This makes MI share traits with open source development. As such, we claim that MI is suited towards utilizing an open experiments platform, as it is not a discipline that relies heavily on closed academic or industry structures. Given that the validities of MI practices are already frequently discussed in these online venues (Casper, 2023), organizing these discussions in a user-friendly platform with added functionality is a natural and helpful extension.

This platform design partly resembles ARBOR, an open collaborations project by Bau Labs (a leading MI lab) in which users can post MI research questions for reasoning models and find collaborators, sharing partial results on experiment pages.[5] Our platform differs in that it is focused on study verification and guideline development. It also shares principles with the MI method leaderboards made by Mueller et al. (2025b).

Educating a Wider Audience. Another benefit from clear guidelines is to make MI more understandable for outsiders and newcomers, allowing academic reviewers who are unfamiliar with the field to use these guidelines as references. Over time, they can potentially evolve and crystallize into canonical textbooks or reference manuals that can be taught in classes at accredited institutes.

Precedents from Established Disciplines. We base the effectiveness of standardized guidelines on precedents from several fields. In biology, MIAME specifies the minimal reporting conventions required for microarray experiments to be interpretable, reproducible, and comparable (Brazma et al., 2001). In clinical practices, the GRADE framework, created by a collaborative of methodologists and clinicians, assesses the quality and certainty of medical studies (Prasad, 2024). In software engineering, High Integrity C++ is a set of programming guidelines for enforcing reliable, maintainable C++ code in automotive, aerospace, and embedded systems (Basalaj and Corden, 2013).

3.2 Principles for Guideline Creation

To ensure guideline quality, the community would benefit from developing meta-guidelines on how to create guidelines. We provide examples as follows:

Guideline Structure. The structure of a *guideline* is flexible, ranging from a short statement to guides with multiple pages.

Minimal Guidelines. We propose that these guidelines should not enforce overly rigorous and unjustifiably strict standards. Rather, we advocate that they should only clearly define the minimal requirements that an experiment should adhere to, and allow flexibility in other areas. Personal "arbitrary" preferences that are not strongly and logically justified should be avoided.

Guides, not Doctrines. We advocate that these

[2] https://www.arena.education/team
[3] http://lesswrong.com/
[4] https://discord.gg/cMr5YqbU4y
[5] https://github.com/ARBORproject/arborproject.github.io

guidelines should not serve as definitive checklists of correctness, but should be used to help practitioners ensure that they are not missing essential elements in their experiments, and to provide more concrete standards for auditors who may not know what to look for.

Encourage Evolution. We expect that an audit system will take time to mature. It should not be adopted or trusted until it has been tested and refined in practice. Even after refinement, the guidelines will be subject to change.

4 A Collaborative Platform for Continuous Reviewing

4.1 A Collaborative Meta-Analysis Platform

We propose establishing a community-driven platform that allows users to post any result, claim, or hypothesis which others can comment on and continue. This approach differs from research archives (arXiv; Papers with Code; Figshare; F1000Research) and journals that collect reproductions (Yildiz et al., 2020) or negative results (Journal of Interesting Negative Results) as it aims to cultivate a community to continuously interact with and check results. It may be viewed as extending LessWrong with added functionalities, described in this Section and Appendix A (with examples of useful meta-analysis comments in Appendix A.1). As such a platform would not be as useful without active engagement, we call for a community effort to encourage participation.

Motivation. Many *meta-analysis* discussions which take place outside of a paper are highly useful for verifying existing work. These discussions and post-hoc results (e.g., replications, reproductions, small counter-results) contain **meta-knowledge** about papers, including small comments and results which do not fit the paper requirements, but are still highly valuable. We discuss these with more detail in Appendix A.1. We suggest that this meta-analysis is a form of continuous reviewing, and already exists in blog posts, Twitter threads, forum posts, private correspondances between researchers, and more. However, its meta-knowledge is often buried, scattered, and not dispersed efficiently to others in a community, staying within those who follow or are in correspondence with researchers who can share meta-knowledge. For instance, replies on Discord may be lost within a server, and Twitter threads, forum posts, and servers are at risk of deletion. Meta-knowledge is also often disorganized and hard to parse. For instance, Twitter threads are not optimally designed for scientific discussion, and their content is questionable if those who control them are compromised. These factors make it inefficient to learn good practices for those who are new to the rapidly growing MI community, and LLMs, which may not have been trained on this information, may also be less efficient at finding and using it.

Therefore, conducting meta-analysis, along with archiving meta-knowledge, in a centralized platform with organizational tools (filtering, recommending, etc.) increases its practical usability. We contend that continuously refining meta-knowledge is essential for developing auditing guidelines, as described in Section 4.3. Additionally, we propose that meta-knowledge is a type of **institutional memory**, which is important community knowledge that is often inadequately documented, and is at-risk of being lost. Previous work has argued that it is essential to record insitutional memory (Corbett et al., 2018).

We posit two essential components to optimally leverage meta-knowledge: *(1) Organization to improve information sharing* and *(2) Live Community-Efforts*. Our view is that this platform should support two main applications:

Use 1: Continuous, Live Reviewing. This is complementary to peer review as this platform is not formal enough to serve as official reviewing; rather, it assists in revision. We propose an approach similar to OpenReview with verified experts and anonymous reviewing, with details in Appendix A.3.

Use 2: Decentralized Collaboration. This platform would allow researchers to not just comment on claims and results, but to share any experiments (partial, post-hoc, etc.) uploaded in **experiment repositories** with collaborative features similar to GitHub. These repositories host hypotheses, evidence, and claims, along with links to code and papers.[6] Researchers can comment on repositories, and collaborate to tweak ideas and verify details. They can continue partial results that others stopped, were stuck on, or requested help on, due to reasons such as a lack of time. Through iterative contributions, multiple repositories may gradually bring about peer-reviewed papers, and the platform tracks all the (known) contributors to specific results, incentivizing platform engagement.

[6] Similar to https://huggingface.co/papers/trending, but with many more reviewing-centered functionalities.

4.2 Encouraging Incentives to Engage in Meta-Analysis

While some researchers engage in meta-analysis discussions, reviews, and research, we contend that there currently is a lack of strong incentive for others to engage in this practice. Given that it can take time, effort, costs, and resources to conduct meta-analysis, and authors are under deadlines to optimally maximize publishing novel papers, engaging in "cleaning work" would comparatively result in little prestige, community engagement, or career-building outcomes, especially on checking "small parts of papers". However, we argue that because such work influences auditing practices, it is important to strengthen incentives to increase engagement, and claim that this platform has the potential to create several strong, new incentives:

1. **Building a Reviewer Portfolio:** This platform may provide user-profile driven incentives for reviewing, such as reviewers who strive to be "expert reviewers" in certain areas, similar to users on Stack Overflow[7]. Users can review any work, mainly those that are relevant to them, and build a portfolio of reviews. One incentive is that this can act as a resume if it becomes an acceptable standard, similar a GitHub portfolio. The reviews shown publically on a profile are the non-anonymized ones, or those that are later non-anonymized. For non-anonymized posts, a user can click on a profile to see an user's works, which includes experiments they have run, their personal views on claims, papers, and more.

2. **Building a Meta-Analysis Portfolio:** Having a platform to share any result may incentivize researchers to undertake useful projects that would not result in a paper, such as negative results that suggest not taking a certain approach (Karl et al., 2024). As the platform would have features that facilitate partial contibutions, it may encourage more non-paper/partial results to be seen and continued.

3. **Becoming a Partial Contributor:** This platform can utilize a contributor assignment system similar to GitHub, such as showing reproductions and replications a researcher contributed to, negative results they pursued that did not result in a paper, and small corrections to a work they assisted with.

[7]https://stackoverflow.com/questions

Currently, there are incentives for users on a platform such as LessWrong to engage in online meta-analysis discussion, as these blog posts and comments are shared within the community. However, there is still room for improvement on how this meta-knowledge is utilized. We see this positive engagement on LessWrong as a precedent supporting the view that community-based meta-analysis for MI is practical and valuable, and suggest that organizing these discussions into a platform more optimized for meta-analysis research will increase this engagement even further, especially for newcomers or those who are not part of the community. For instance, there can be pages dedicated to certain claims (e.g., the Linear Representation Hypothesis) which collects links to these discussions, and more importantly, encourages continued engagement with a much wider audience via features such as recommendation systems on this particular topic, sharing them as "trending topics" that capture more attention.

Incentives Under Anonymity. This experimental approach can test various anonymity options to assess their effectiveness and flaws. It is possible to still improve in meta-analysis engagement users who are fully anonymous (with no link to a profile) or under a pseudonym (an profile that is not linked to public identity). For instance, this platform can provide a wider and more engaging community that they otherwise would not find on other platforms, and it can provide organizational tools, recommendation systems that suggest work they would be interested in, and more which improve their meta-analysis and reviewing experience, even if the community is fully or partially anonymous.

4.3 How a Collaborative Platform Establishes Auditing Guidelines and Protocols

Both professional auditing and reviewing platforms can falter if its users are not aware of good reviewing practices. Thus, it is crucial to crystallize meta-knowledge about standardized reviewing practices, guidelines, and protocols. We propose a discussion system in which communities work together to generalize common, useful and expert-supported points developed on this platform into guidelines to improve auditing efficiency, both in professional, regulatory practices and on the platform. We suggest for this platform to allow researchers to create *Proposed Guideline* pages where they debate with others over their validity, justifying arguments with empirical sources.

By organizing and connecting experiments together in a shared space, this platform can help the community identify and refine good practices by looking for common patterns across studies, combining their insights into practical auditing guidelines which inform governance and regulation. Additionally, waiting for official “next editions" of guidelines is a slow process, while MI is a rapidly evolving field; hosting “living document" guidelines that can be quickly updated can accelerate innovation.

Figure 2 shows an example of how these guideline discussion pages are formatted, with users posting evidence-based arguments supported by community-reviewed experiment repositories. We describe stages on how these freeform discussions lead to standardized, adopted guidelines:

1. Users suggest *Proposed Guidelines* pages with arguments supported by empirical evidence, citing papers, repositories, and more. For instance, a Guideline page may propose "Use Test X" because previous discussions by researchers found that running Test X from Study A was highly important to assess circuit validity.

2. On each *Proposed Guideline* page, the community debates about the guideline’s validity. For instance, researchers may be against the guideline due to crucial flaws in Study A, as discovered by a small meta-result. This may be further supported by Study B findings.

3. When a professional auditing system seeks to select standardized guidelines, they can consult the arguments on the Proposed Guideline page as evidence-backed references. This ensures that these important meta-analysis discussions are not lost, organizing them in one place. If the auditing system is to be updated, they can track updates via the platform.

We provide high-level examples of guidelines in Appendix C. To assist with organzed cross-study comparisons, these experiment formats may benefit from being built on a unified protocol language. We suggest that the community work towards building a user and model friendly protocol which includes formalizing hypothesis testing of causal explanations, and describe this in Appendix F.

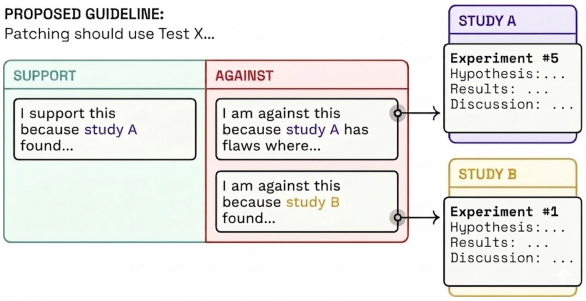

Figure 2: Example for how proposed guideline discussion pages organize arguments supported by community-reviewed experiment repositories.

5 Source-Based Automated Auditing

Source-Based Reasoning. To assist with uncovering claim validity, we suggest developing tools which trace the assumptions, evidence, and other claims which claims depend on. This means that claims discussed on the platform are scrutinized more carefully based on the other statements they depend on, allowing one to trace back statements and evaluate them recursively. This is more specific than citations as this does not just reference papers, but the specific statements within them, such plots. We discuss this further in Appendix A.4.

Automated Auditing Assistance via Agentic AI. One issue is that there are too many claims for manual source assessment. Thus, we propose the development of automated systems to assess claims at scale that aim to be more objective, possibly using the guidelines curated by the MI community. This approach becomes more important as more automated interpretability methods are developed to accompany human-driven analyses. This automated system may be built with agentic systems that help trace back long dependency chains of claims, while humans manually verify their outputs to check for and fix issues like hallucinations. These agentic systems may also run the code from papers and experiment repositories via evaluation harnesses. Additionally, this system may benefit auditors by converting explanations into testable claims with explicit priors that an independent party can recompute.

Scoring Claims via Probabilistic Logic. To assist with how reasons affect claims, researchers and AI agents may find benefits with using probabilistic logic to weigh claims based on claims assumptions (Appendix A.4 that discusses this using "Claim Views"). This provides quantifiable estimates of a claim’s validity, assisting researchers in

their personal understanding of each claim's certainty. However, we note that such a system only suggests values for humans to ultimately verify.

Overall, this system weighs the strength of a claim by the reasons it depends on to be valid. For instance, if a claim previously depended on assumption X, and assumption X was found to be false, then the claim is automatically updated to be very dubious, and a warning alert is sent to those who follow this claim.

We propose an automatic evidence-weighing auditing system be built using rigorous logical probabilistic verification frameworks such as Probabilistic Soft Logic (PSL) (Kimmig et al., 2012), in which hypotheses and observations are linked by logical relations, and weighted rules can be automatically updated such that the likelihood of evidence is maximized, reducing the burden on researchers to hand-tune assumptions. This system could support handling audit questions such as whether hypotheses were preregistered or post hoc, how sensitive conclusions are to ablations or counterfactuals, exposing when conclusions rest on selective evidence, settling competing explanations, and mitigating the influence of implicit biases. In Appendix G, we discuss the advantages of systems like PSL for auditing, and give an example of a use case for PSL on auditing a study claiming circuit minimality. We also illustrate a high-level example of automated auditing using a simple LLM based auditing system in Appendix H.

6 Alternative Views and Considerations

We describe two alternate views below and discuss three more in Appendix D which address: (i) Guideline definition difficulty, (ii) Being too early to define standards, and (iii) If the field is evolving too fast to have standards.

View 1: Why verify using defined guidelines? Aren't practical outcomes sufficient? One concern is that practical effectiveness is sufficient evidence of correctness. For instance, if a model steers well, then it "works". However, we still have to define what "steering well" means, as without these more rigorous definitions, it is subject to subjective human interpretation. Essentially, these guidelines aim to ensure that experimental conclusions mirror real-world effectiveness as best as possible. Steering well should be defined based on outcomes of recorded empirical data. It should be noted if past steering experiments conducted on a certain benchmark were representative enough after those approaches were deployed in safety-critical contexts. If it was found that they were not, and that this benchmark rested on brittle assumptions or missed untested edge cases, it is paramount, as a guideline, to re-test the approach on additional benchmarks that better reflect real-world distributions.

View 2: Standardization is restrictive and discourages adoption. Another concern is that formal standards could stifle creativity or impose arbitrary requirements that do not generalize across different research agendas. If researchers do not immediately perceive value in such guidelines, they will be reluctant to adopt such a framework.

However, we do not propose to develop guidelines which are unjustified and strict, but rather, are lightweight, minimal, and necessary enough to just cover what is needed for high quality experiments. Each guideline should be logically and empirically justified; if a standard is considered arbitrary and not generalizable, the community should encourage its removal. This leaves substantial room for creative methodologies and novel hypotheses.

Given that these guidelines aim to be highly useful in strengthening the validity of competing claims, we expect that researchers would see value and demand in this framework.

7 Conclusion

Mechanistic interpretability has produced valuable insights, but without standardized auditing procedures the field lacks reliability in safety-critical deployment and governance. Regulatory frameworks increasingly mandate behavioral transparency, and post-hoc explainability methods can greatly improve truth in AI used in financial services, healthcare, and insurance sectors. We advocate for the reseach community to pursue the development of continuous reviewing, a collaborative approach that gradually refines pieces of research using *meta-analysis* results and discussions that fit outside of a paper. We propose that this development will benefit via improved organization of meta-research work and discussions using community-driven platforms. Establishing auditing guidelines arising from community-driven meta-research that gradually critiques and refines work outside of papers can greatly assist in refining interpretability for trustworthy AI.

Limitations

Our position advocates for a goal to establish standardized empirical guidelines for MI, which requires substantial community coordination. We posit an experimental reviewing approach via a platform that depends on active user engagement, which we have not gathered yet, but aim to test in future work, possibly via surveys and workshops. This position paper seeks to obtain this user engagement by encouraging researchers to share their opinions on the necessity and effectiveness of our proposed direction for the MI community, and on how well it can produce standardized guidelines for auditing. Without catalyzing these discussions, there would be fewer users and less attention given to testing this experimental approach in initial trials. Additionally, for real-world guideline adoption by regulatory bodies, key challenges include defining governance structures for guideline development, balancing minimal requirements with methodological flexibility, establishing enforcement mechanisms without stifling innovation, and scaling verification to frontier models where exhaustive testing becomes computationally prohibitive. We do not claim to resolve these tensions fully, as this paper aims to catalyze discussion, rather than to impose solutions.

References

Armaan A. Abraham. 2025. Deep sparse autoencoders yield interpretable features too. https://www.lesswrong.com/posts/tLCBJn3NcSNzi5xng/deep-sparse-autoencoders-yield-interpretable-features-too.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9525–9536, Red Hook, NY, USA. Curran Associates Inc.

Dario Amodei. 2025. The urgency of interpretability. https://www.darioamodei.com/post/the-urgency-of-interpretability.

Anthropic. 2025. Claude sonnet 4-5 system card. https://assets.anthropic.com/m/12f214efcc2f457a/original/Claude-Sonnet-4-5-System-Card.pdf.

arXiv. arxiv e-print archive.

Daniel Balsam, Myra Deng, Nam Nguyen, Liv Gorton, Thariq Shihipar, Eric Ho, and Thomas McGrath. 2024. Goodfire ember: Scaling interpretability for frontier model alignment. https://www.goodfire.ai/blog/announcing-goodfire-ember.

Wojciech Basalaj and Richard Corden. 2013. High integrity c++ coding standard, version 4.0. Whitepaper, Programming Research Ltd.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and advances. *Preprint*, arXiv:2102.12452.

Leonard Bereska and Stratis Gavves. 2024. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*. Survey Certification, Expert Certification.

Alvis Brazma, Pascal Hingamp, John Quackenbush, Gavin Sherlock, Paul Spellman, Chris Stoeckert, John Aach, Wilhelm Ansorge, Catherine A. Ball, Helen C. Causton, Terry Gaasterland, Patrick Glenisson, P. Holstege Frank C. Irene F. Kim, Victor Markowitz, John C. Matese, Helen Parkinson, Alan Robinson, Ugis Sarkans, and 5 others. 2001. Minimum information about a microarray experiment (miame) — toward standards for microarray data. *Nature Genetics*, 29(4):365–371.

Stephen Casper. 2023. The engineer's interpretability sequence. Alignment Forum (series of posts).

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. 2022. Causal scrubbing: a method for rigorously testing interpretability hypotheses. https://www.lesswrong.com/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. 2025. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *Preprint*, arXiv:2409.14507.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. 2023. A toy model of universality: reverse engineering how networks learn group operations. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Consumer Financial Protection Bureau. 2023. Consumer financial protection circular 2023-03: Adverse action notification requirements and the proper use of the cfpb's sample forms provided in regulation b.

Consumer Financial Protection Bureau. 2025. Supervisory highlights: Advanced technologies special edition, issue 38 (winter 2025). Winter 2025 Edition.

Jack Corbett, Dennis C. Grube, Heather Lovell, and Rodney Scott. 2018. Singular memory or institutional memories? toward a dynamic approach. *Governance*, 31(3):555–573.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition.

European Parliament and Council of the European Union. 2024. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act). Published July 12, 2024.

F1000Research. F1000research: An open research publishing platform.

Lucy Farnik. 2025. Great to see so much discussion of our recent paper! leo's take here is likely true. https://x.com/lucyfarnik/status/1885967412375290097.

Figshare. Figshare: Research data repository.

Dan Friedman, Andrew Lampinen, Lucas Dixon, Danqi Chen, and Asma Ghandeharioun. 2024. Interpretability illusions in the generalization of simplified models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Roman Gansch, Lina Putze, Tjark Koopmann, Jan Reich, and Christian Neurohr. 2025. Causal bayesian networks for data-driven safety analysis of complex systems. *Preprint*, arXiv:2505.19860.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. 2025. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Preprint*, arXiv:2301.04709.

Ashkan Golgoon, Khashayar Filom, and Arjun Ravi Kannan. 2024. Mechanistic interpretability of large language models with applications to the financial services industry. In *Proceedings of the 5th ACM International Conference on AI in Finance*, ICAIF '24, page 660–668. ACM.

Ryan Greenblatt. 2023. How useful is mechanistic interpretability? https://www.lesswrong.com/posts/tEPHGZAb63dfq2v8n/how-useful-is-mechanistic-interpretability.

Rohan Gupta, Iván Arcuschin, Thomas Kwa, and Adrià Garriga-Alonso. 2024. Interpbench: Semi-synthetic transformers for evaluating mechanistic interpretability techniques. *Preprint*, arXiv:2407.14494.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. 2025. Sparse autoencoders can interpret randomly initialized transformers. *arXiv e-prints*, pages arXiv–2501.

Journal of Interesting Negative Results. Journal of interesting negative results.

Florian Karl, Lukas Malte Kemeter, Gabriel Dax, and Paulina Sierak. 2024. Position: Embracing negative results in machine learning. *Preprint*, arXiv:2406.03980.

Angelika Kimmig, Stephen Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2012. A short introduction to probabilistic soft logic. In *Proceedings of the NIPS workshop on probabilistic programming: foundations and applications*, pages 1–4.

Georg Lange, Alex Makelov, and Neel Nanda. 2023. An interpretability illusion for activation patching of arbitrary subspaces. *AI Alignment Forum*. https://www.lesswrong.com/posts/RFtkRXHebkwxygDe2/an-interpretability-illusion-for-activation-patching-of.

David Lindner, Janos Kramar, Sebastian Farquhar, Matthew Rahtz, Tom McGrath, and Vladimir Mikulik. 2023. Tracr: Compiled transformers as a laboratory for interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pages 37876–37899. Curran Associates, Inc.

Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. 2018. Deepproblog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Samuel Marks. 2025. In fact, this experiment has been done at least 3 times! https://x.com/saprmarks/status/1885869042629857590.

Samuel Marks, Johannes Treutlein, Trenton Bricken, Jack Lindsey, Jonathan Marcus, Siddharth Mishra-Sharma, Daniel Ziegler, Emmanuel Ameisen, Joshua Batson, Tim Belonax, Samuel R. Bowman, Shan Carter, Brian Chen, Hoagy Cunningham, Carson Denison, Florian Dietz, Satvik Golechha, Akbir Khan, Jan Kirchner, and 16 others. 2025. Auditing language models for hidden objectives. *Preprint*, arXiv:2503.10965.

Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. 2023. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Raphaël Millière and Charles Rathkopf. 2026. Anthropocentric bias in language model evaluation. *Computational Linguistics*, 52(1):379–388.

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. 2025a. The Quest for the Right

Mediator: Surveying Mechanistic Interpretability for NLP Through the Lens of Causal Mediation Analysis. *Computational Linguistics*, pages 1–48. _eprint: https://direct.mit.edu/coli/article-pdf/doi/10.1162/COLI.a.572/2554934/coli.a.572.pdf.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, and 4 others. 2025b. MIB: A mechanistic interpretability benchmark. In *Forty-second International Conference on Machine Learning*.

Maxime Méloux, Giada Dirupo, François Portet, and Maxime Peyrard. 2025a. The dead salmons of ai interpretability. *Preprint*, arXiv:2512.18792.

Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. 2025b. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? *Preprint*, arXiv:2502.20914.

Jatin Nainani, Sankaran Vaidyanathan, AJ Yeung, Kartik Gupta, and David Jensen. 2024. Adaptive circuit behavior and generalization in mechanistic interpretability. *arXiv preprint arXiv:2411.16105*.

Neel Nanda. 2025. How to become a mechanistic interpretability researcher. AI Alignment Forum.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*.

Oscar Obeso, Andy Arditi, Javier Ferrando, Joshua Freeman, Cameron Holmes, and Neel Nanda. 2025. Real-time detection of hallucinated entities in long-form generation. *Preprint*, arXiv:2509.03531.

Theo X Olausson, Alex Gu, Benjamin Lipkin, Cedegao E Zhang, Armando Solar-Lezama, Joshua B Tenenbaum, and Roger Levy. 2023. Linc: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *arXiv preprint arXiv:2305.12295*.

Papers with Code. Papers with code.

Manya Prasad. 2024. Introduction to the grade tool for rating certainty in evidence and recommendations. *Clinical Epidemiology and Global Health*, 25:101484.

Matthew Richardson and Pedro Domingos. 2006. Markov logic networks. *Mach. Learn.*, 62(1–2):107–136.

Tim Rocktäschel and Sebastian Riedel. 2017. End-to-end differentiable proving. *Preprint*, arXiv:1705.11040.

Naomi Saphra and Sarah Wiegreffe. 2024. Mechanistic? In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 480–498, Miami, Florida, US. Association for Computational Linguistics.

Second Look Research. Second look research. `https://secondlookresearch.com/`.

Harald Semmelrock, Tony Ross-Hellauer, Simone Kopeinik, Dieter Theiler, Armin Haberl, Stefan Thalmann, and Dominik Kowald. 2025. Reproducibility in machine-learning-based research: Overview, barriers, and drivers. *AI Magazine*, 46(2):e70002.

Glenn Shafer. 1976. A mathematical theory of evidence.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeffrey Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Isaac Bloom, Stella Biderman, Adrià Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Mary Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, William Saunders, and 10 others. 2025. Open problems in mechanistic interpretability. *Transactions on Machine Learning Research*. Survey Certification.

Claudia Shi, Nicolas Beltran-Velez, Achille Nazaret, Carolina Zheng, Adrià Garriga-Alonso, Andrew Jesson, Maggie Makar, and David M. Blei. 2025. Hypothesis testing the circuit hypothesis in llms. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

David Soergel, Adam Saunders, and Andrew McCallum. 2013. Open scholarship and peer review: a time for experimentation.

Xiangchen Song, Aashiq Muhamed, Yujia Zheng, Lingjing Kong, Zeyu Tang, Mona T Diab, Virginia Smith, and Kun Zhang. 2025. Position: Mechanistic interpretability should prioritize feature consistency in saes. *arXiv preprint arXiv:2505.20254*.

Dashiell Stander, Qinan Yu, Honglu Fan, and Stella Biderman. 2024. Grokking group multiplication with cosets. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Denis Sutter, Julian Minder, Thomas Hofmann, and Tiago Pimentel. 2025. The non-linear representation dilemma: Is causal abstraction enough for mechanistic interpretability? *Preprint*, arXiv:2507.08802.

Daniel Tan, David Chanin, Aengus Lynch, Dimitrios Kanoulas, Brooks Paige, Adria Garriga-Alonso, and Robert Kirk. 2025. Analyzing the generalization and reliability of steering vectors. *Preprint*, arXiv:2407.12404.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. 2024. LLM circuit analyses are consistent across training and scale. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482.

U.S. Food and Drug Administration. 2025. Artificial intelligence-enabled device software functions: Lifecycle management and marketing submission recommendations - draft guidance for industry and fda staff. Draft Guidance, Docket No. FDA-2024-D-4488.

U.S. Food and Drug Administration and Health Canada and Medicines and Healthcare products Regulatory Agency. 2021. Good machine learning practice for medical device development: Guiding principles.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Rowan Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Some lessons learned from studying indirect object identification in gpt-2 small. https://www.lesswrong.com/posts/3ecs6duLmTfyra3Gp/some-lessons-learned-from-studying-indirect-object.

Ke Weng, Lun Du, Sirui Li, Wangyue Lu, Haozhe Sun, Hengyu Liu, and Tiancheng Zhang. 2025. Autoformalization in the era of large language models: A survey. *arXiv preprint arXiv:2505.23486*.

John Wu, David Wu, and Jimeng Sun. 2025a. Dila: Dictionary label attention for mechanistic interpretability in high-dimensional multi-label medical coding prediction. In *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, pages 1014–1038. PMLR.

Wilson Wu, Louis Jaburi, Jacob Drori, and Jason Gross. 2025b. Towards a unified and verified understanding of group-operation networks. *Preprint*, arXiv:2410.07476.

Burak Yildiz, Hayley Hung, Jesse H. Krijthe, Cynthia C. S. Liem, Marco Loog, Gosia Migut, Frans Oliehoek, Annibale Panichella, Przemyslaw Pawelczak, Stjepan Picek, Mathijs de Weerdt, and Jan van Gemert. 2020. Reproducedpapers.org: Openly teaching and structuring machine learning reproducibility. *arXiv preprint arXiv:2012.01172*.

Fred Zhang and Neel Nanda. 2024. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*.

Appendix

A A Continuous Reviewing Platform for Meta-Results: Extended Description

A.1 How do we Define Meta-Results?

Meta-results contain information that is useful, but often is not found within a paper. These include critiques, comments (e.g., small comments on one part of a paper, even on an appendix study), seemingly minor negative results, replications (conducting the exact same study), reproductions (conducting the study under varied conditions), non-novel extensions (e.g. just add an ablation study, try it in a new framework, etc.), minor results (which, by themselves, are not enough for a paper, but can be useful in future if built on by others), and partial results (e.g., still requires an ablation study).

Experiment repositories allow hosting findings that are fall outside papers, such as a specific feature/circuit found using existing methods which are not novel enough for a paper, but can useful enough to share, potentially contributing to auditing by finding counter use-cases, or compiled into a database with other minor findings.

Examples of useful meta-results can be found in comments on platforms such as LessWrong (Wang et al., 2022; Chan et al., 2022; Lange et al., 2023; Greenblatt, 2023; Abraham, 2025) and Twitter/X (Farnik, 2025; Marks, 2025).

For instance, users have posted meta-results in the form of post-hoc re-evaluations or reproductions, such as a comment in a LessWrong post where a researcher revisited IOI (Wang et al., 2023) using causal scrubbing (Chan et al., 2022) and found that it held up differently than the original paper's faithfulness-style metric suggested, with recovered effects that was only fractions of the original logit differences (Wang et al., 2022). They state that there is "uncertainty in interpreting these numbers"; we posit that small, specific open questions like these are useful for passing along to others who can help contribute to them, and propose tools that give them greater visibility. Additionally, these questions may have likely been answered later (e.g., this is an older topic that already had a good amount of discussion), and so archiving how they were answered later is also valuable.

Another example of valuable meta-analysis is a discussion over claims from a paper on randomized SAEs, in which researchers defended their claims, clarified misunderstandings, and shared meta-results (Farnik, 2025; Marks, 2025).

However, we note that not all meta-analysis discussions and results are as neatly archived as these LessWrong discussions. In addition, it would be helpful to have a place to showcase and find these results better, such as using organized listings and networks of experiment repositories related to certain claims (e.g., the Linear Representation Hypothesis), as this can improve collaborative meta-analysis. Storing meta-results in experiment repositories, instead of just reporting them in comments that are response to blogs, can enhance reproducibility, their use in future work, and their use in informing guidelines.

A.2 Building a Collaborative Reviewing Culture

To pass down meta-knowledge efficiently, it would be advantageous to cultivate a reviewing culture that encourages sharing information and good practices. A "Review Hub" platform can allow users to highlight to make comments, like Overleaf or Google Docs, and ask questions if something is unclear. This platform also helps each user personally understand claims better and share their understanding with others. Continuous reviewing and commenting would also allow targetting very specific parts of a paper or experiment (e.g., Why is this plot justified? Why use this metric?), rather than requiring users to review entire papers.

Such a platform would not serve as official peer review, given that it is less robust to issues such as bias and quality, but would serve as a drafting and brainstorming tool that assists with collaboratively verifying and improving upon claims. It can also help assess a researcher's own personal beliefs. Thus, the platform does not aim to assign scores to projects/papers. Optionally, the platform may have systems which give scores to claims, but these are merely suggestions used in one factor of validation. This helps drafts before they are put in peer review.

A.3 Filtering based on Expert Opinions

One issue with a public system that any user can comment on is that there can be too much information, including spam. Thus, the platform can benefit users by providing them with multiple, customizable options for what information to filter (e.g., hide content with many down votes, show all content without voting influence, etc.) Having multiple options can allow users to assess multiple viewpoints, rather than just locking into potentially biased communities or authoritative figures, while also assessing selective expert opinions to find con-

tent that is assumed to be more likely of higher quality. On project pages, users and authors may filter to only show content from 'expert-verified' profiles. Verification can be done similar to how it is vetted on OpenReview, or may directly import an OpenReview profile. Though the profile is verified, the identity can be anonymized to the public. Double blind anonymity or non-anonymity options may be allowed.

A.4 Tracking Assumptions via Claim Dependency Graph Tools

We propose a system to improve how researchers assess argument validity via "claim dependency lists/graphs". These tools may help with collaborative debates by formatting, organizing, and visualizing large information traces better. We define **statements** as pieces of content: phrases, entire papers, a plot, etc. We define **claims** as statements asserted to be true (to some degree). We propose a tool in which, given a claim, users construct arguments by linking to pieces of other claims, ending at root statements (assumptions) that are subjectively satisfactory for the current debate.

During debates, a researcher can ask to further unfurl a claim (e.g., ask "what does it depend on?"), and also debate what reasons each statement should actually depend on. One user may propose one reason to support a claim, and another may accept it, or question that reason even further, asking for more scrunity on a certain benchmark it claimed to have passed.

Essentially, creating a dependency tracking tool allows researchers to keep asking for sources of sources while keeping track of potentially long dependency chains, with each claim node having with multiple proposed reasons supporting them. This may help with recursively questioning assumptions that claims are dependent on, which can become unorganized as one keeps on tracing reasons back. For instance, a researcher may assume result X from paper P is true. When constructing an argument, researchers may have unquestionably assumed this was true as they did not explicitly unfurl this reason before. However, with this tool, they have be more prone to questioning it, and finding "bugs" they missed before.

Due to the many reasons that claims depend on, and dependency chains which may go on infinitely, no dependency graph is assumed to be exhaustive. During arguments, a researcher's claim depends on certain key reasons; the dependency graph only needs to be made to contextually be useful in that specific debate. They can refer to specific areas of some piece (e.g., a specific plot). Then, reviewers hone in on that plot and debate the veracity of that plot, since its validity contributes to the claim.

Thus, a claim graph is not exhaustive; it is an organizational scratchpad tool used to aid in discussions over claim validity. As such, it only displays a select number of claims which a claim depends on in the given context. Given one important claim which the paragraph is centered on defending, we create a graph of claims which the claim depends on. From there, we can further unfurl each claim node into possible reasons to support them. During discussions, a researcher can further add reasons as to why a reason supports a claim, and mark the edge as "weak" or "strong".

We note that claim graphs cannot replace written arguments, as the reasons which support a claim can often be too complex to fully capture in a claim graph. However, they are only meant as an organizational tool to assist in thinking and discussion.

Implementation Details. Claims are created by users. Researchers can state "I create this claim and show what it depends on". Users can click on a claim to open a UI element which shows what that claim depends on. They can keep on clicking on claims to open up sources of sources. This can be shown as a list or graph visualizer, with the former being easier to use for claims with many dependencies.

Users can propose a claim in a paper/experiments/claim/blog/guideline page and give what reasons they believe it depends on. In public pages, users vote to keep what proposed claims should be shown. Others can subjectively state that a claim still needs more reasons. In the public views, a researcher can see the votes agreeing with this or not (with the option of not having votes in personal views unless turned on by the host). Claims or dependencies that are voted down by a well-calculated metric are hidden or removed to prevent clutter.

The platform would aim to allow an easy-to-use construction of claim graphs. Claim graphs can be saved onto the platform to assist others in understanding the reasons supporting a claim, and allowing them to suggest improvements, such as by questioning assumptions (e.g., this metric was found to have a flaw). A researcher can trace back claims of claims to find more relevant pieces to review.

Claim Pages. Optionally, some claims can become pages. Researchers can create pages that focus on claims themselves, rather than on specific papers. This means researchers focus on the results, and less on authors of specific papers. This allows results from multiple papers to be collected into a more organized and archived discussion, allowing arguments that support claims to be traced more specifically to parts of papers, not just citing entire papers.

Automated Claim Weighing. Claims can be assigned validity scores based on other claims they depend on. A researcher assigns weight to them; by default, one has assumed statements that are true. Note that this is often not objective enough, but just "convincing enough". Hence, we do not have strong conclusive "weighing systems"; the weighing systems are only meant to give suggested, estimated, calculated snapshot scores, similar to statements made by LLMs. It is up to humans to use those scores or not.

A.4.1 Claim Dependency Views

Given that the reasons which claims are dependent upon can be subjective, we suggest the use of Claim Dependency Views (i.e., **Views**), which are subjective dependency graphs of what claims depend on. Different researchers may disagree on what reasons a claim depends on, and thus can build their own views to share with other researchers, which may improve information organization and ease of collaborative debates. Users can fork views to continue in their own way, and historically trace where they were built off from.

To prevent spam by bots and trolls, there can be are "verified" views which anyone who is verified can comment on. This helps prevent gatekeeping from experts, but has the downside that of noisier arguments from non-experts. Thus, to filter out the opinions of non-experts, there is can be a "verified expert" View. Verification is done by ensuring a user is a human using identification systems. Experts are verified similar to how conferences select for reviewers, based on reputation and references. In Verified Expert views, experts can only interact in the field or sub-field they are verified in. Others can propose changes to a view, which the author can accept. To prevent clutter, public views can have reasons removed by a sophisticated voting consensus (to be determined) which is designed to be robust against manipulation.

Overall, we propose four main types of public Views: General, Verified, and Verified Expert, along with Personal Views. Within each view, one can show them in three different ways: All displays every post without votes, Vote-Filtered displays with Votes, hiding those with very low votes to prevent clutter, and Custom is user-customized (e.g., filter out certain users). For personal views, a user can create as many copies as they want to help with validity reasoning, forming different conclusions under different assumptions.

Each argument is measured by several metrics, including community votes (both unweighted and reputation-weighted metrics). Users can rank arguments in various ways by filtering for certain metrics. Guidelines may be scored across multiple metrics, including argument certainty (based on validity of prior assumptions, etc.)

A.5 Benefits of Decentralized Collaboration

Researchers can post work that needs to be done or checked, laying out proposed roadmap for others to take up. Other users can fork and continue repos in their own way. This helps prevents partial progress or questions from being forgotten, and helps maintain existing work with minor extensions.

A.6 Recommendation and Search Systems

To fuel community engagement and direct researchers to interact with work they are familiar with, this platform can benefit by using a recommendation system to relay and relate results. Users can customize what is recommended, and choose claims, experiments, and drafts to review. They can search for and filter by expertize, sub-field, and specific claims they tackle.

A.7 Organizational Tools for Cross-Experiment Analysis

Currently, experiment results are recorded in papers or blog posts. These results are enshrouded in dense, unstructured text; it would be beneficial to distill them into more structured formats for easier comparison. Organized repositories allow users to more easily search, filter, and cross-compare fine-grained results across studies. Users can search for specific experiments and their variations (e.g., hyperparameter sweeps) within a study, checking their thoroughness. Users can create custom dashboards from which they can cross-compare results, and notice patterns that they can abstract into general guidelines.

B A Step-by-Step Example of Practical MI Auditing

B.1 Example of Auditing a Study for Safety-Critical Medical Scenarios

Setting. Clinicians want to detect signs of a heart attack from short patient descriptions. They use a model to determine whether patients have signs of a heart attack or not.

Use of MI for fine-tuning / pruning. When clinical guidelines change (e.g., atypical presentations in women, diabetics, or elderly patients), knowing the internal circuit lets developers fine tune or prune only the relevant components instead of retraining the entire model. Hospitals can also re-audit over time or across subpopulations.

Experiment Claims. After activation patching, the researchers conclude that specific model components perform the reasoning needed to flag high-risk cases. Their paper presents highly impressive results.

Step by Step Audit. We describe an audit using activation patching guidelines.

1. *Frame the objective.* Auditors define the target behavior precisely. The model must label a short description as urgent when symptoms like chest pressure and radiating arm pain appear.

2. *Collect Reproducibility Artifacts.* They pin the checkpoint, tokenizer, seeds, hardware, code commit, and the exact evaluation set. This ensures that any divergence is meaningful rather than accidental.

3. *Read the data.* They inspect a sample of clean prompts and the study's corrupted counterparts. The corrupted texts replace key symptom phrases with realistic alternatives that alter clinical meaning while preserving grammar.

4. *Test the corruption.* They verify that corruption produces a measurable drop in correct urgent labeling, yet keeps text fluent and in distribution. This confirms that a restoration effect would matter.

5. *Probe granularity.* They retrace the authors' path from residual stream to layer to head to token position. Patch effects appear where the paper reports them, which suggests the localization procedure was applied carefully.

6. *Establish baselines.* They run and archive three states. Clean, corrupted, and patched corrupted. The corrupted state shows degraded performance and the patched state seems to restore it.

7. *Small scale reproduction.* Using fresh random seeds and equivalent prompts, they reproduce headline plots on a held out slice. Effect sizes are similar, which increases confidence in implementation fidelity.

8. *Critical metric audit.* The auditors notice a dire issue: **the paper scores success only by the probability of the urgent label**. Auditors recompute with *logit difference*, comparing urgent against a non urgent alternative,

$$\Delta\ell = z(\text{urgent}) - z(\text{non urgent}).$$

   Several components that looked important under probability show negligible or even negative $\Delta\ell$. The probability gains came from incidental shifts in the output distribution, not from genuine causal influence on the decision.

9. *Decision and remediation.* Since the metric masks the true effect, **the central claim does not hold**, and the study cannot be used yet for clinical decision-making. The study must be repeated with logit based or KL based evaluations, pre registered metrics, and the same otherwise strong design. Corruption, granularity, and baselines were sound, but the metric choice invalidates the causal conclusion.

Note that just because a study passes one auditing test, does not mean it is automatically robust. For one, there may be aspects to test which that auditing test is missing. However, it does mean that the study is seen as more robust than before the auditing was conducted. Therefore, auditing ascribes more certainty to the study's validity, increasing confidence in its use in safety-critical applications.

B.2 Example of Generalized Auditing Steps for Activation Patching

1. Define the target behavior and hypotheses.

2. Collect all artifacts: model version, code, prompts, random seeds, and metrics.

3. Validate prompt construction and the evaluation set for relevance and leakage.

4. Examine the corruption method; confirm it is in-distribution and causes a measurable behavioral shift.

5. Verify that the evaluation metric aligns with the hypothesis (e.g., logit difference rather than raw probability).

6. Confirm that the patch targets and granularity match the causal claim.

7. Establish clean and corrupted baselines prior to patching.

8. Reproduce a subset of patch results to confirm the reported effect.

9. Perform sensitivity analyses with alternative metrics, corruptions, seeds, and prompt distributions.

10. Review causal interpretation, check for negative or redundant components, and document limitations and recommendations.

On the online platform, users will be able to propose freeform guides such as this example. The community can then discuss, agree with, and critique this item, refining it over time.

C More Guideline Examples

Examples of Guidelines Types.

1. Guidelines for hypothesis testing

2. Guidelines for evaluating observations (i.e., do the study's claims match what the methods actually find?)

3. Guidelines for comparing methodologies

4. Guidelines for designing benchmarks

Examples of Standardized Definitions (for Circuit Discovery): We base these defintions on the theoretical framework of causal abstraction (Geiger et al., 2025).

- **Hypothesis**: a high-level causal model H over low-level model internals N that posits how inputs, latents, and outputs relate.

- **Feature**: a component in the high-level model H that carries a specific causal role.

- **Causal Abstraction**: a mapping τ that relates low-level variables in N to high-level variables in H, preserving (or approximately preserving) causal relationships under permissible interventions.

- **Faithfulness**: a quantitative measure of how closely the high-level model H approximates the interventional behavior of N.

Examples of Criteria to Test (for Circuit Discovery):

1. **Behavior Preservation**: Intervening to route model computation through the proposed circuit must preserve the model's task behavior relative to the unmodified model within a small, predefined tolerance.

2. **Localization**: Perturbations restricted to the hypothesized circuit should reproduce the effect, and perturbations outside the circuit should not.

3. **Minimality**: Remove components that are not necessary without reducing performance on the target behavior beyond a preset tolerance.

4. **Distribution Shift Robustness**: Guarantee that their interpretations can be maintained across distribution shifts in the data.

Examples of Benchmark Standards:

- Test on intervention benchmarks like MIB (Mueller et al., 2025b) and InterpBench. (Gupta et al., 2024) to compare circuit localization and causal variable localization across methods. For instance, MIB provides a benchmark with standardized tasks, counterfactual datasets, principled metrics, and private-test-set leaderboards for both localization and featurization, encouraging more stable cross-method comparison. InterpBench complements this by providing 86 semi-synthetic transformers with known circuits, making the ground-truth causal structure more trustworthy, so researchers can test whether a method actually recovers the right mechanism rather than only producing plausible explanations.

- Include stress tests across i.i.d. and out-of-distribution shifts to assess robustness of interpretability claims.

Pitfall	Description	Auditing Guideline
Statistical fragility (Méloux et al., 2025a)	Plausible-looking explanations may not be uniquely identified by the evidence.	Test alternative hypotheses and quantify uncertainty across null baselines and perturbed settings
Over-reliance on automated interpretability (Heap et al., 2025)	Automated labeling methods may produce seemingly interpretable features even in randomly initialized models, suggesting that labels may not correspond to genuine concepts.	Validate features through causal tests and comparison with random or null models
Feature instability across SAE runs (Song et al., 2025)	Sparse autoencoders trained with different random seeds can produce different features, making single-run interpretations unreliable.	Check whether features are consistently recovered across multiple SAE training runs
Ignoring redundant or adaptive circuits (McGrath et al., 2023)	Networks may contain redundant pathways that compensate for ablations, as shown by the Hydra effect, making single-pathway explanations incomplete.	Test robustness of explanations under multiple ablations and intervention patterns
Anthropomorphic projection (Millière and Rathkopf, 2026)	Human-intuitive narratives may be projected onto model internals without reflecting the model's actual computation.	Prefer operational or causal descriptions over anthropomorphic narratives
Sensitivity to prompt format or dataset curation. (Nainani et al., 2024)	Circuit behavior can change substantially with prompt format, data distribution, or random seed choices.	Evaluate findings across multiple prompt templates, datasets, and random seeds

Table 3: More examples of potential pitfalls in MI experiments, and guidelines to audit whether experiments avoid them. Auditing checks if these guidelines are followed.

- Enforce replicability checks by requiring that published experiments be rerun with small perturbations (e.g., input noise or weight initialization variation).

Examples of Model Organism Standards:

- Declare reference models that are tractable to inspect.
- Require version control of published checkpoints, hyperparameters, and training logs to ensure reproducibility.
- Standardize documentation of dataset provenance, training protocol, random seeds, architecture details.

Examples of General Practices:

- **Reporting standard:** Report intervention operators, ablation knobs, thresholds, datasets, and code used to run the tests. Use a structured schema so an auditor can recompute the same statistics.
- **Ground-truth validation where possible:** When a ground-truth mechanism exists, require that methods recover it before application to opaque models. Use Tracr compiled transformers as "known-mechanism" testbeds to validate tooling and to calibrate localization and minimality metrics (Lindner et al., 2023).
- **Benchmark against community tasks:** Evaluate methods on public, multi-task benchmarks that score both circuit localization and causal variable localization.
- **Sanity checks and falsification attempts:** Run randomization and model-parameter permutation test to ensure results are model and

data-dependent. If a purported circuit explanation survives when labels or weights are randomized, the explanation fails a minimum bar. Include negative controls and counterexamples by design.

- **Leakage and benchmark contamination checks:** Verify that no information from evaluation tasks, hidden test sets, or benchmark-specific artifacts can enter training, augmentation, hyperparameter tuning, or prompt design; when leakage is plausible, confirm results on stricter holdout or semi-private splits and disclose all benchmark-exposure pathways.

Examples of Hypothesis Formation Advice:

- **Beware Bias for Elegant Algorithms:** Models may not prefer simple, elegant and human-understandable algorithms.
- **Models are lazy learners:** Models prioritise prediction accuracy over deep understanding. If there is a simple heuristic that allows accurate predictions, they will likely learn that and cease learning. As long as they get the right answer, they are done.
- **A model algorithm sub-task tends to improve prediction accuracy:** A model only learns a subtask if it improves prediction accuracy - even if just a little in an edge case. This can give you clues on what sub-tasks might exist. e.g. in Addition, a sub-task to map pairs of digits (e.g. "2 with 4 maps to 6", termed BaseAdd) will give the correct prediction if there is no carry-one from the previous column. So for some questions, it gives the correct answer. That makes it useful enough and the model learns it.

Examples of Auditing Red Flags:

- **Purely qualitative visuals or cherry-picked cases without falsification:** Adebayo et al. (2018) show that visually compelling explanations can be independent of model or data. Any claim that relies on visual appeal without randomization checks is insufficient.
- **Subjective human scoring as the primary validator:** Reviews emphasize that human evaluation is inconsistent and non-scalable. Use automated, preregistered tests and report uncertainty.
- **Uncalibrated localization:** If perturbations outside the circuit also change behavior at similar magnitudes, the claim lacks specificity (identify true negative circuits correctly).
- **Benchmarks omitted or misinterpreted:** Ignoring community baselines leads to overclaimed novelty.

Examples of Experiment Goals:

- **Goal A:** Comparable claims across labs. Express every claim with the same primitives: a set of nodes, edges, interventions, and quantitative outcomes for preservation, localization, and minimality. Provide machine-readable artifacts so that automated verifiers can ingest results and recompute scores.
- **Goal B:** Generalization beyond toy demos. Require that every confirmed claim passes the three property tests on out-of-distribution probes of the capability.

D More Alternative Views

We continue addressing alternative views from Section §6 in this section.

View 3: Defining these guidelines will be difficult

A similar counterargument to #2 is that there are tasks in mechanistic interpretability that are difficult to assess objectively and precisely. For instance, whether steering works well or not may be subject to arbitrary, case-specific thresholds that do not generalize.

Mirroring our previous response, we propose that these guidelines will be written not to define every assessment, but to ensure that the majority of mechanistic interpretability studies follow the best practices built on expert consensus that has been shaped by feedback from real-world applications. As such, they will avoid introducing unjustified rules, such as claiming that steering can only work well according to exact thresholds, if historical studies demonstrate that these thresholds depend on a specific case-by-case basis that are difficult and cumbersome to quantify.

Instead, these guidelines will ensure minimal practices are followed, and if they are not or could not be (such as having barriers like being unable to apply a certain diagnostic method due to technical limitations), they will estimate the degree of uncertainty in hypothesized claims.

For steering “well", this means not forcing its definition to be within an arbitrary threshold, but to ensure that steering experiments are conducted on benchmarks that are representative of real-world scenarios for the target task (e.g., coding). Approaches do not are marked as “largely uncertain".

View 4: It is too early to introduce standards

This argument states that the field is too young for standardization, and that premature formalization could restrict methods or discourage exploration. However, the intent of this proposal is not to impose a finished framework immediately. Instead, we call for a structured, collaborative process to begin developing one that starts with open discussion, evolving drafts, and iterative refinement.

It is precisely because the field is growing rapidly that initiating this effort early is advantageous. Early attempts will not be perfect, but they will surface areas of disagreement, reveal practical needs, and help shape a community record of arguments, revisions, and precedents. As AI capabilities advance, delaying the creation of verification infrastructure increases the risk that interpretability research will become too vast and heterogeneous to organize retroactively.

View 5: The field is evolving too fast for the existing techniques to be standardized

New techniques in MI are constantly being developed, and it can be argued that unlike in established practices such as medicine, these techniques will frequently be replaced such that developing standards for one technique will be useless as that technique will not be useful very quickly.

We propose that there are general approaches in MI which stand the test of time, and are unlikely to go away, such as activation patching, steering, and circuit discovery. Guidelines developed for these general approaches will also likely remain steadfast. As an analogy, many new software tools are constantly being developed, but general “good coding and SWE practices" have endured.

E Importance of MI Auditing for Regulatory Policies

Regulations impose outcome-oriented transparency requirements while remaining agnostic to implementation methods. The EU AI Act (European Parliament and Council of the European Union, 2024) mandates transparency sufficient for “deployers to interpret a system’s output and use it appropriately" (Article 13). However, accompanying guidance acknowledges the regulation “does not clearly address the actual level of transparency required." Violations carry fines up to €35 million or 7% of global annual turnover, whichever is larger.

Consumer Financial Protection Bureau (2023) articulates the most technically specific requirements, and rejects generic adverse action explanations, demanding that credit denials based on “behavioral spending data" identify which specific behaviors triggered the decision. Consumer Financial Protection Bureau (2025) escalated the requirements: examiners now actively search for Less Discriminatory Alternatives using open-source tools, scrutinize models with over 1,000 variables for proxy discrimination, and require rigorous validation of adverse action methodologies.

The FDA’s evolution illustrates regulatory maturation without mechanistic specificity. The most comprehensive FDA AI document to date (U.S. Food and Drug Administration, 2025) emphasizes “clear interpretation mechanisms" among its ten Good Machine Learning Practice principles (U.S. Food and Drug Administration and Health Canada and Medicines and Healthcare products Regulatory Agency, 2021) while acknowledging that authorized devices exhibit “often limited explainability of AI predictions".

F MI Protocols

A shared protocol format allows smoother application of guidelines to audit specific instances. For instance, if a guideline says "measure minimality of feature X to test hypothesis H", knowing what specifically should be defined as feature X, and how to measure minimality, would be beneficial and reduce confusion.

For scalability and to mitigate human errors, experiments can include a protocol that defines how an automated AI system translates unstructured results into structured formats. This protocol allows humans to work together with automated interpretability systems, to assess discoveries at a scale that manual efforts cannot. If each result is then expressed in the same schema, automated comparison and meta-analysis become feasible, turning a patchwork field into a coherent, cumulative discipline. We expect that such a protocol would not be immediately built, but will crystallize over time as guidelines become more established. An example of the layers of this protocol is given in Table 4.

Protocol Layer	Purpose	Details
Community Rules (Human Layer)	Define norms, requirements, and minimum validity criteria	A living "standards document" (e.g. web-wiki), community review processes, versioning, rule proposals & voting
Abstract Framework / Schema	Translate rules into a canonical formal language. Each author can define a schema	Schema definitions (e.g. JSON), "MI hypothesis" primitives, required test definitions, metric templates
Machine-Readable Protocol	Encode the framework so tools can parse, check, and verify claims	APIs, validation scripts, automatic verifiers, benchmark suite

Table 4: An example of abstraction layers in a MI protocol

G Automated Auditing Assistance using Probabilistic Logic

An automatic auditing system would parse the evidence provided for a hypothesized claim and suggest how certain that claim is, while identifying components that are incomplete or improperly executed. The protocol would format experimental data into a structured representation suitable for logical verification and iterative confidence estimation.

Logical frameworks such as PSL can help filter out noise and curb the hallucinations that may arise in LLM-driven auditing (Pan et al., 2023; Trinh et al., 2024). We propose combining the two approaches: large language models would specialize in parsing and structuring unstructured experimental result data, while logical probabilistic frameworks such as PSL would act as verification layers that translate this parsed evidence into quantitative, internally consistent estimates of certainty. Together, this hybrid design would allow automatic auditing systems to interpret natural-language experiment descriptions and evaluate them through a rigorous, rule-based uncertainty model.

Recent progress in neurosymbolic reasoning provides a foundation for this approach. Systems such as *Logic-LM*, which integrates symbolic solvers to improve the faithfulness of LLM reasoning (Pan et al., 2023), and *LINC*, which couples first-order logic provers with language models for verifiable reasoning (Olausson et al., 2023), demonstrate that hybrid reasoning pipelines can produce both interpretability and formal soundness. Parallel efforts in *autoformalization* (Weng et al., 2025), which aim to convert unstructured scientific or mathematical outputs into logical representations, further highlight the feasibility of this integration by bridging free-form LLM output with symbolic verifiers.

A particularly relevant precedent is the recent *Nature* study on *Solving Olympiad Geometry without Human Demonstrations* (Trinh et al., 2024), which showed that coupling LLMs with structured geometric solvers yields precise, verifiable reasoning in a complex symbolic domain. This work illustrates how a similar strategy could underpin automated auditing: by combining LLMs for linguistic and evidential parsing with probabilistic logic systems for formal evaluation, one can achieve both expressive understanding and rigorous verification of experimental claims.

Example of using PSL to Assist with Automated Auditing. In PSL, predicates (representing properties or relationships) take on soft truth values in [0,1], allowing degrees of truth rather than binary outcomes. Given experimental data, PSL infers predicate values and learns rule weights by optimizing a convex objective, aggregating disparate pieces of evidence into calibrated confidence.

We demonstrate an example with a sample set of predicates for a circuit minimality problem. Assuming we have some explicit performance metrics, we break down the problem into observed variables, and latent variables.

Grade Mapping Predicates. We define a set of "grade" predicates for the Minimality score:

Predicate	Condition
Min_ge_095(C)	Minimal$(C) \geq 0.95$
Min_ge_090(C)	Minimal$(C) \geq 0.90$
⋮	⋮
Min_ge_070(C)	Minimal$(C) \geq 0.70$
Min_ge_060(C)	Minimal$(C) \geq 0.60$

1. **InCircuit(E,C)**: Edge E is in candidate cir-

cuit C (binary variable).

2. **Sufficient(C)**: Circuit C meets the behavioral or specification threshold (binary).

3. **Cost(C,V)**: Normalized cost or size of C in $[0, 1]$ (real-valued).

4. **PerfDrop(E,C,V)**: Normalized performance drop in $[0, 1]$ when removing E from C (higher = worse performance).

5. **RemMass(C)**: Precomputed or aggregated removable mass for C in $[0, 1]$ (e.g., the sum or average of Removable(E, C) over all edges).

Latent Targets. These are the unobserved variables that must be inferred by the system:

1. **Critical(E,C)**: (*latent*) Edge E is necessary for the sufficiency of C (degree in $[0, 1]$).

2. **Removable(E,C)**: (*latent*) Removing E preserves the sufficiency of C (degree in $[0, 1]$).

3. **Minimal(C)**: (*latent*) Circuit C is approximately minimal with respect to cost, given sufficiency (degree in $[0, 1]$).

Grade Labels (latent)

Predicate
Grade_A(C)
Grade_Aminus(C)
⋮
Grade_D(C)
Grade_E(C)

This provides a general recipe adaptable to other use cases. There are other alternative probabilistic and logic based systems, such as Dempster-Shafer Theory (Shafer, 1976), that may apply to such systems, as described in Table 5. We encourage the community to develop and experimentally compare such variants across interpretability use cases, gradually identifying which frameworks best balance interpretability, scalability, and epistemic rigor in automated auditing.

H Auditing Minimal-Circuit Claims

Mechanistic interpretability papers frequently present explanations in the form of *minimal circuit stories*: a small set of heads, neurons, or features that are claimed to implement a task. However, such explanations can be misleading if key validity checks are missing. Inspired by recent discussions on auditing MI claims, we propose a simple checklist that evaluates whether a circuit explanation satisfies basic evidentiary standards.

To operationalize this idea, we implement a lightweight auditing tool that parses a written explanation and checks whether it contains evidence for a set of minimal-circuit guidelines. The tool prompts an LLM to evaluate whether the explanation provides explicit evidence for each guideline.

Audit Prompt

You are evaluating a mechanistic interpretability explanation that claims (or might claim) to have found a minimal circuit for some task.
For each guideline below, determine whether the explanation provides clear evidence that the guideline is satisfied. If evidence is missing or unclear, mark it as false. Guidelines include: sufficiency evidence, necessity evidence, sparsity of the identified circuit, robustness on held-out data, causal validation through interventions, sanity checks against null baselines, and stability across random seeds or pruning tie-breaks.

Figure 6 shows the output of the auditing tool on a toy Indirect Object Identification (IOI) circuit explanation. The example satisfies most minimal-circuit criteria but fails checks for multiple initializations and tie exploration during greedy pruning.

This checklist is not meant to be exhaustive, but it provides a practical framework for identifying common fallacies in circuit explanations. In particular, missing seed robustness or pruning tie exploration can indicate that the reported circuit may be unstable or one of several equally plausible solutions.

Table 5: Representative probabilistic and logic-based systems and how they might tackle circuit minimality auditing.

Framework	Application to Circuit Minimality
Probabilistic Soft Logic (PSL) (Kimmig et al., 2012)	Represents sufficiency and edge importance as soft predicates inferred jointly, quantifying uncertainty about minimality under conflicting evidence.
Markov Logic Networks (MLNs) (Richardson and Domingos, 2006)	Encodes weighted logical rules such as "if edge E is critical, removing it should reduce performance," estimating the probability that a candidate circuit is minimal given observed interventions.
DeepProbLog (Manhaeve et al., 2018)	Combines symbolic rules for minimality (e.g., necessity and sufficiency) with neural submodules that estimate probabilistic truth values from activation representations.
Neural Theorem Provers (NTPs) (Rocktäschel and Riedel, 2017)	Differentiably matches logical patterns against learned embeddings of activation traces to infer relations such as "edge E contributes to function F."
Causal Bayesian Networks (CBNs) (Gansch et al., 2025)	Represents directional dependencies between components and outputs, allowing interventions (e.g., "remove E") to estimate causal effects and test minimality.

Minimal-Circuit Guideline	Status
Claims minimal/sparse circuit	✓
Sufficiency evidence	✓
Necessity evidence	✓
Sparsity / small circuit	✓
Robustness (held-out / shifts)	✓
Non-spurious (not cherry-picked)	✓
Causal validation (patching / ablations)	✓
Sanity checks / null baselines	✓
Multiple initializations / seeds	✗
Tie exploration in greedy pruning	✗

Table 6: Example scorecard produced by the minimal-circuit auditing tool. The explanation provides strong causal and robustness evidence but does not analyze stability across seeds or tie-breaking choices in greedy pruning.