# White Box Evidence Packages for Policy Audit Reports

**Seunghyun Yoo** [1]

## Abstract

As AI governance moves from benchmark scores toward auditable oversight, a central question is how reviewers can tell whether an LLM generated audit report is actually supported by evidence. This paper studies that question in passage anchored policy audits, where a report must interpret a given policy passage and cite evidence for its claims. We introduce a controlled evaluation framework that holds the passage, rubric, and auditor model fixed while changing only the evidence interface supplied to the auditor. Across 60 AGORA policy cases, we generate 600 structured reports under ten evidence conditions, including passage based evidence, internal model evidence, a hybrid package, and a shuffled control that preserves evidence format while breaking case relevance. Five human reviewers evaluate the primary interfaces for correctness, passage grounding, diagnostic usefulness, and evidence misuse. The results show that internal evidence changes how reports cite and reason about evidence, but more internal citations do not by themselves make a report more valid. A white box diagnostic explains the failure mode: causal localization is narrow, while reports readily reuse broader readable labels and token directions. The hybrid interface is the most useful on average, while the shuffled control exposes a key governance risk: reports can sound substantively plausible while citing irrelevant internal evidence. This study reframes internal model access as an evidence design problem for audit workflows, rather than as a guarantee of transparency.

[1]Governance and Responsible AI Lab. Correspondence to: Seunghyun Yoo <yoo236@purdue.edu>.

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

## 1. Introduction

AI governance increasingly needs audit reports that are not only persuasive, but checkable. Regulatory audits, AI safety institute evaluations, DSA style platform audits, and assurance standards all push oversight toward claims that can be traced to evidence rather than only to model outputs or benchmark scores (International Auditing and Assurance Standards Board, 2013; Solarova et al., 2026; Holznagel, 2025; Reuel et al., 2024b). Black box access remains essential because it captures what external users can observe, but it does not by itself answer a harder question: whether an audit report used the right evidence for the claim it makes (Casper et al., 2024). This paper studies that evidence problem for LLM assisted policy audit reports.

We focus on passage anchored policy audits. In this setting, a report must interpret a given policy passage, cite supporting evidence, and make uncertainty visible to a later reviewer. The passage is the source of truth; the report is not a legal compliance decision, a human auditor experiment, or a capability audit where model internals may be the primary object of study. The key experimental question is simple: if the passage, rubric, and auditor model stay fixed, how does the evidence interface change the report?

The evidence interfaces compare surface evidence, internal model evidence, hybrid evidence, and a relevance broken control. Here, white box evidence means auditor facing entries produced from fixed interpretability tools applied to Gemma 2 2B. These entries are not treated as trusted explanations. They include text, source information, and caveats that the auditor may cite in a structured report. The report is therefore the unit of analysis: its findings, passage spans, evidence citations, confidence, and misuse of evidence.

Figure 1 summarizes the design. We construct 60 AGORA policy cases, generate 600 structured reports with a local Qwen 2.5 7B auditor across ten evidence conditions, and conduct human gold and diagnostic review with five reviewers. The primary human review covers four interfaces over all 60 cases: a strong black box surface baseline, a full white box package, a hybrid package, and a shuffled white box control that keeps the evidence format while pairing internal evidence with the wrong case.

The results support three claims. First, internal evidence

changes citation behavior, but citation uptake is not evidence validity. Second, hybrid packages are the most useful interface in this design, but their advantage is a design signal rather than a causal proof because the hybrid package is larger and more varied. Third, the failure mode is shallow use rather than non use: the auditor repeatedly cites readable SAE labels and token direction summaries, while a residual patching diagnostic localizes the strongest causal signal to a narrow layer window. The contribution is therefore not a claim that white box access generally improves audits. It is a reproducible framework for evaluating internal model evidence as part of a reviewable audit workflow. To support reproducibility, we provide code, prompts, evidence packages, final reports, human review files, and paper figures at https://anonymous.4open.science/r/agora-mi-C473/README.md.

## 2. Related Work

Audit access work motivates the need for evidence interfaces, but access is not the same as usable evidence. Technical AI governance work argues that oversight institutions need tools for access, assessment, verification, and accountability rather than only post hoc descriptions of model outputs (Reuel et al., 2024b;a; Ojewale et al., 2025). Related audit access work argues that black box interaction can miss information needed to test developer claims and system risks, motivating controlled access to weights, activations, development evidence, or verification channels (Casper et al., 2024; Waiwitlikhit et al., 2024; Harack et al., 2025). Our study accepts that motivation but evaluates the next step: whether internal information can be packaged, cited, and checked in an audit report.

Audit practice makes evidence relevance central. Assurance standards and regulatory audit regimes ask auditors to connect conclusions to sufficient and appropriate evidence, while early DSA audit critiques show that formal audit reports can remain weak when audit criteria, evidence relevance, and systemic risk tests are underspecified (International Auditing and Assurance Standards Board, 2013; Solarova et al., 2026; Holznagel, 2025). Access scholarship similarly warns that API and platform restrictions can create audit blind spots even when transparency mandates exist (Burnat & Davidson, 2025; Cen & Alur, 2024). Stakeholder first transparency work argues that explanations should be tailored to the decision maker and use context rather than treated as generic disclosures (Bell et al., 2022). We apply these ideas to LLM generated audit reports: an evidence package should make evidence available, but it should also make relevance and misuse reviewable.

Mechanistic interpretability provides candidate sources of internal evidence, but readable tool outputs still need downstream validation. Sparse autoencoders can turn residual stream activations into feature hypotheses (Cunningham et al., 2024; Bricken et al., 2023; Templeton et al., 2024; Lieberum et al., 2024); logit lens methods can summarize intermediate token directions (Belrose et al., 2023); steering and sensitivity methods can test how internal directions affect model behavior (Cyberey et al., 2026); and automated explanation methods can convert activations into natural language hypotheses (Bills et al., 2023; Karvonen et al., 2025a). Recent evaluation work warns that readable labels and tool outputs should be tested through downstream use, ablations, and controls rather than treated as explanations by default (Karvonen et al., 2025b; Sheshadri et al., 2026). AuditBench is closest in spirit because it evaluates multiple white box tools as audit instruments; our setting differs by evaluating the downstream policy report artifact while holding the same passage and rubric fixed.

The shuffled control draws on known LLM context reliability failures. Prior work shows that language models can be distracted by irrelevant context, fail to use long context reliably, over rely on retrieved passages, and exhibit sycophantic agreement patterns under social or evidential pressure (Shi et al., 2023; Liu et al., 2024; Yoran et al., 2024; Sharma et al., 2023). We do not claim to newly discover that LLMs can overuse irrelevant evidence. Our contribution is to instantiate that risk in governance facing white box evidence packages and measure whether reports rely on internal evidence that belongs to the wrong case.

## 3. Audit Task and Evidence Interfaces

Each case asks the auditor to produce a reviewable report from a policy passage. The input contains a policy passage, a fixed audit rubric, and one evidence package. The hidden gold brief is withheld from the auditor and used only for human review. The output report must identify policy relevant issues, cite supplied evidence entries and exact passage spans, record uncertainty, and avoid claims not supported by the package.

This task fixes the authority structure of the audit. The passage is the object being interpreted, and white box evidence is auxiliary evidence about how a target model represents or reacts to that passage. This differs from a capability or behavioral audit where there may be no passage level ground truth and internal evidence may be the main object of investigation. Our results should therefore be read as evidence about LLM assisted passage grounded policy audit reports, not as a ranking of white box tools for all audit settings.

The experimental control is deliberately simple: the passage, rubric, and auditor stay fixed, and only the evidence package changes. Every package contains stable evidence entries, source information, text, and caveats. The auditor is asked to cite an entry only when it directly supports a finding. This

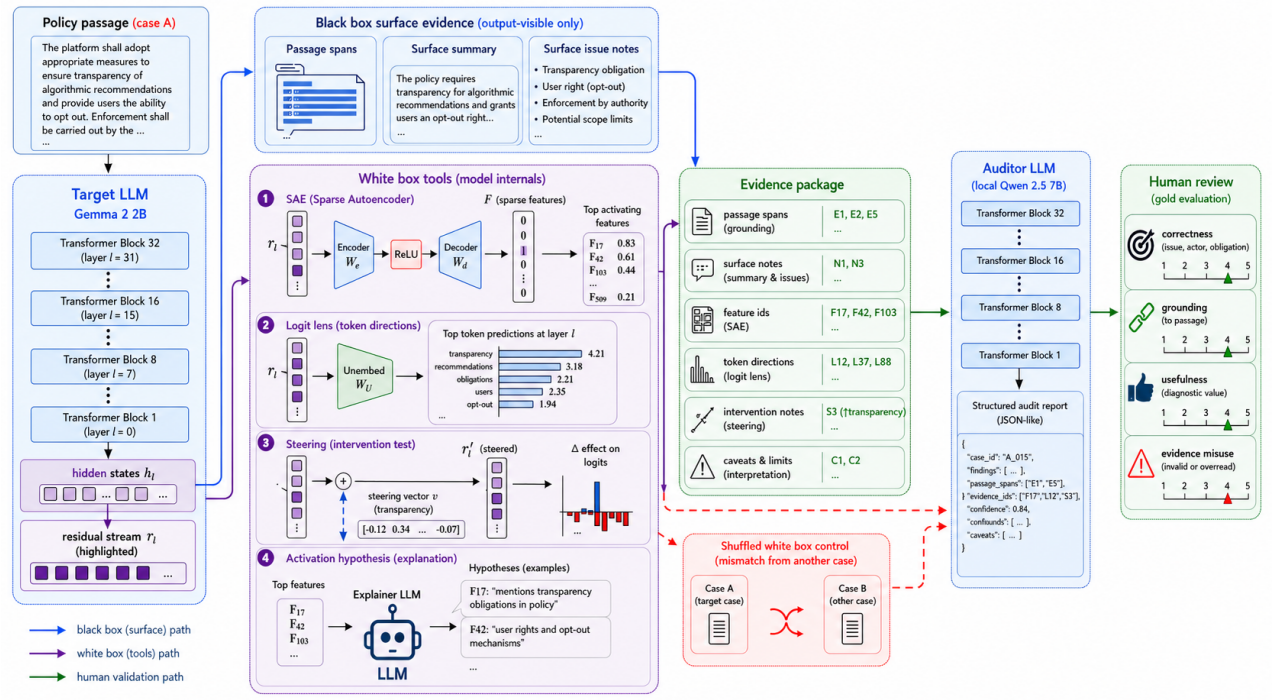


*Figure 1.* Evidence package audit design. Policy passages are processed by a target model, internal states are summarized through SAE, logit lens, steering, and activation explanation surrogate records, and the resulting evidence package is given to an auditor model. The unit of comparison is the evidence interface, not a retrieval system. Human gold briefs are written from the policy passage before diagnostic review of auditor reports.

design separates three objects that audit reports often blur: the policy passage, surface observations about the passage, and internal evidence about the target model.

Table 1 lists the ten evidence conditions. The main narrative focuses on four primary interfaces. C1 is a strong black box surface baseline built from selected passage spans, issue notes, and confidence annotations. C6 combines all available white box views. C7 adds the C6 internal evidence to the C1 surface anchors. C9 keeps the shape of the C6 package but draws internal evidence from another case; it is the relevance control that tests whether plausible looking internal evidence is treated as authoritative when it should not be.

The remaining conditions are diagnostic single tool interfaces. They expose SAE features, logit lens token directions, steering sensitivity records, activation explanation surrogates, or raw AutoInterp labels in isolation. These conditions help interpret tool behavior, but the main quality claim comes from the four primary interfaces that receive full human review. Tool construction details and formulas are provided in the appendix.

*Table 1.* Evidence conditions.

| ID | Interface | Auditor facing contents |
|---|---|---|
| C0 | Passage | Passage and rubric only. |
| C1 | Black box | Surface summary, selected spans, issue notes, and confidence. |
| C2 | SAE | Feature activations, layers, labels, spans, and caveats. |
| C3 | Logit lens | Intermediate token direction summaries. |
| C4 | Steering | Directional sensitivity and output or score deltas. |
| C5 | Surrogate | Natural language activation hypotheses. |
| C6 | Full white box | C2 through C5 together. |
| C7 | Hybrid | C1 plus C6. |
| C8 | Raw SAE | Raw AutoInterp labels and feature references. |
| C9 | Shuffled | C6 style evidence drawn from another case. |

## 4. Experimental Setup

**Cases and models.** We use 60 policy cases drawn from AGORA, a corpus of AI governance and regulatory material (Arnold et al., 2024). The cases come from 11 source documents and include obligations, rights, exceptions, enforcement mechanisms, and governance procedures. The interpreted target model is Gemma 2 2B, and the auditor is a local Qwen 2.5 7B model run with structured outputs. For each condition, the auditor receives only the passage, rubric, and assigned evidence package. It does not see the hidden gold brief and is not told that C9 contains shuffled evidence.

**Evidence generation.** The package generator is deterministic given the selected case set. SAE items use Gemma Scope residual stream SAEs and human revised AutoInterp labels where available. Logit lens items use Gemma 2 2B layers 6, 12, 18, and 24 at audit cue token positions selected from the passage. Steering items use layers 12, 18, and 24 with a fixed obligation and risk contrast prompt set. The activation explanation condition is a conservative surrogate derived from selected SAE evidence and label information; it tests the interface role of natural language activation hypotheses, not a validated vector decoder.

**Report normalization.** Every response is normalized into a structured report with findings, passage spans, cited evidence, internal evidence use, confidence, and cautionary notes. Structural checks verify that each report can be read, contains findings, records confidence consistently, and cites only evidence entries that were actually supplied. These checks establish that the pipeline produced inspectable reports, not that the reports are accurate.

**Human review.** Human review has two stages, and five reviewers contribute across gold writing, verification, diagnostic scoring, and agreement checks. First, reviewers write blind gold briefs from the policy passage and source information before seeing model reports or evidence packages. Second, reviewers score selected reports against those briefs. The primary review covers all 60 cases under C1, C6, C7, and C9, yielding 240 scored reports for correctness, span grounding, diagnostic usefulness, and evidence misuse. Agreement diagnostics use the original 20 case subset, and optional single tool diagnostic review covers C2, C3, C4, C5, and C8 on that subset. Lower is better only for misuse; the appendix gives the full review protocol.

Scores are summarized as condition means over the reviewed case set. Pairwise statements compare each condition with C1 on the same cases and report better, tied, and worse counts; for misuse, lower scores are better. We use exact sign tests and Wilcoxon signed rank approximations as descriptive paired checks because the scores are ordinal and tied. Citation uptake is the mean number of evidence citations per report, and the shuffled warning counts C9 cases with both high correctness and high evidence misuse. Full notation and score distributions are in the appendix.

*Table 2.* Human review axes. Scores use a 1 to 5 scale. Higher is better except for misuse.

| Axis | Review question |
|---|---|
| Correctness | Does the report match the gold issue, actor, and obligation structure? |
| Span grounding | Are major claims tied to concrete passage text? |
| Usefulness | Would the report help an auditor notice the issue or gap? |
| Evidence misuse | Does the report overread, overcite, or rely on irrelevant internal evidence? |

## 5. Results

The results answer four report quality questions and one mechanism question: whether internal evidence changes evidence use, whether it improves human reviewed report quality, whether a hybrid interface helps, which white box evidence the auditor actually cites, and whether relevance controls catch plausible but invalid evidence use.

**RQ1: Does internal evidence change evidence uptake?** Yes. The Qwen 2.5 7B auditor produced reports for all 600 condition and case combinations, and all reports were successfully normalized after retrying long responses. Structurally, full white box evidence changes the report interface substantially: C6 reports cite 10.13 evidence entries per report versus 3.87 under C1. However, C9 shuffled white box reports also cite 9.65 evidence entries per report. Citation volume is therefore evidence uptake, not evidence validity.

**RQ2: Does internal evidence improve reviewed report quality?** Not by itself. The primary human review, conducted with five reviewers across gold brief and diagnostic stages, covers all 60 paired cases under C1, C6, C7, and C9. Table 3 and Figure 2 summarize the results. C6 preserves high average correctness (4.60 vs. 4.68 for C1), but it substantially weakens grounding (3.25 vs. 4.52), lowers usefulness (4.00 vs. 4.68), and increases misuse (2.50 vs. 1.00). The paired pattern is consistent with the averages: C6 is worse than C1 on grounding in 53 of 60 cases and worse on usefulness in 41 of 60 cases, while evidence misuse is higher in all 60 cases.

**RQ3: Does a hybrid interface help?** C7 is the most useful interface on average. It is nearly tied with C1 on correctness (4.70 vs. 4.68) and grounding (4.50 vs. 4.52), while improving usefulness from 4.68 to 4.92. The paired usefulness effect is positive in 17 cases, tied in 40, and negative in 3. However, C7 also increases evidence misuse in all 60 cases, from 1.00 to 2.13 on average. This pattern is consistent with the hypothesis that hybrid evidence can work when surface passage cues anchor the report and internal evidence is used as a secondary signal. It is not a causal isolation of anchoring, because C7 receives a larger package than C1 or C6 on average (22.97 evidence items versus 3.27 and 19.70).

**RQ4: Which white box evidence does the auditor actually cite?** White box evidence is read shallowly rather than ignored, but citation uptake is not mechanistic localization. In raw report citations, C6 cites logit lens evidence most often, with 232 logit lens citations versus 68 SAE citations. After normalizing by how many items are available, however, SAE labels dominate: C6 cites 93.2% of available SAE entries, compared with 33.9% of logit lens entries, 3.1% of

steering entries, and 2.8% of activation surrogate entries. The same pattern survives the shuffled control, where C9 cites 95.9% of available SAE entries despite the evidence coming from another case. A separate residual activation patching diagnostic gives the stricter white box view (Vig et al., 2020; Zhang & Nanda, 2024). Figure 4 shows that deontic force passes the causal gate only at final prompt token layers 7 and 8; nearby layers and other tested regions remain below the empirical threshold. The mechanism level reading is therefore narrower than the report citations: the auditor can reuse readable labels and token directions broadly, while causal support appears only in a small localized window.

**RQ5: Why is a shuffled control necessary?** C9 is the strongest warning. It has lower average correctness than C1 and C7, but it often remains substantively plausible because the auditor can still read the passage. Thirty nine of sixty C9 reports have correctness at least 4 while also receiving evidence misuse at least 4. Thus a report can be mostly right about the policy passage while invalidly citing internal evidence from another case. The auditor's own cautionary notes rarely catch this problem: only 3 of these 39 reports warn about weak evidence, and only one explicitly names an internal tool. Figure 3 shows the same point from another angle: citation uptake cannot be used as a quality measure.

*Table 3.* Primary human review results over 60 paired cases. Scores are means on a 1 to 5 scale. Higher is better except evidence misuse.

| Condition | Correct | Ground | Useful | Misuse |
|---|---|---|---|---|
| C1 black box | 4.68 | 4.52 | 4.68 | 1.00 |
| C6 full white | 4.60 | 3.25 | 4.00 | 2.50 |
| C7 hybrid | 4.70 | 4.50 | 4.92 | 2.13 |
| C9 shuffled | 3.63 | 2.67 | 2.63 | 5.00 |

*Table 4.* Paired deltas relative to C1 over 60 cases. The B/T/W column reports cases where the condition is better than, tied with, or worse than C1 on the axis; for misuse, better means lower misuse, so positive deltas are worse.

| Comparison | Axis | Mean $\Delta$ | B/T/W | Sign $p$ |
|---|---|---|---|---|
| C6 vs. C1 | Correct | $-0.08$ | 8/39/13 | .383 |
| C6 vs. C1 | Ground | $-1.27$ | 0/7/53 | $<$ .001 |
| C6 vs. C1 | Useful | $-0.68$ | 0/19/41 | $<$ .001 |
| C6 vs. C1 | Misuse | $+1.50$ | 0/0/60 | $<$ .001 |
| C7 vs. C1 | Correct | $+0.02$ | 9/43/8 | 1.000 |
| C7 vs. C1 | Useful | $+0.23$ | 17/40/3 | .003 |
| C7 vs. C1 | Misuse | $+1.13$ | 0/0/60 | $<$ .001 |
| C9 vs. C1 | Correct | $-1.05$ | 0/6/54 | $<$ .001 |
| C9 vs. C1 | Misuse | $+4.00$ | 0/0/60 | $<$ .001 |

## 6. Qualitative Evidence From Human Review

The aggregate result is easiest to interpret through actual policy passages. We use the case studies to check whether the numeric pattern means what the paper claims: hybrid evidence can help when passage anchors remain primary, while white box evidence can still mislead when internal labels are treated as substantive policy evidence.

The first case explains why the hybrid interface is useful. Articles 29 and 30 of the Chinese Algorithmic Recommendation Provisions contain two separate issues: confidentiality for supervision bodies and personnel, and a complaint or report mechanism for organizations and individuals. C1 is correct and useful but not fully comprehensive on span grounding (5, 4, 5, 1). C6 finds the related confidentiality theme but weakens grounding and usefulness relative to the passage anchored baseline (5, 3, 4, 2). C7 keeps the passage anchored black box surface evidence visible, so the report covers both confidentiality and complaint handling while using internal evidence secondarily (5, 5, 5, 2). C9 is partly correct but cites shuffled transparency evidence (4, 2, 3, 5). This is consistent with the interface claim: hybrid packages can keep white box evidence tied to the passage, but this case does not isolate anchoring from the larger evidence package.

The second case shows the limitation. The EU AI Act passage is a classification exception: an Annex III system is not considered high risk if it does not pose a significant risk under listed conditions, but profiling of natural persons remains always high risk. C1 captures the not high risk exception and the profiling carve out (4, 4, 4, 1). C6 preserves the broad classification topic but can still blur the legal category and cite weak internal evidence (4, 3, 4, 2). C7 improves anchoring but can still use imprecise low risk language (4, 4, 4, 2). C9 uses shuffled deployment evidence and misses part of the profiling override (3, 2, 2, 5). This is why the paper does not claim that hybrid evidence solves legal precision. It performs best on average, but legal category wording remains fragile.

Across the 39 C9 reports with high correctness and high misuse, we observe three recurring patterns. First, the auditor can read the passage correctly while citing internal evidence from another case. Second, broad SAE or logit lens labels can look relevant to the policy family while failing to support the specific actor, obligation, or exception. Third, shuffled evidence can transfer the report toward plausible transparency, deployment, or risk themes that are not the relevant legal structure. These examples justify the shuffled control as more than a sanity check. A policy audit report can be useful at the surface level and still misuse the internal record. For governance uses, that is precisely the dangerous case: a reviewer who only checks whether the conclusion sounds plausible could miss that the cited internal evidence is irrelevant.

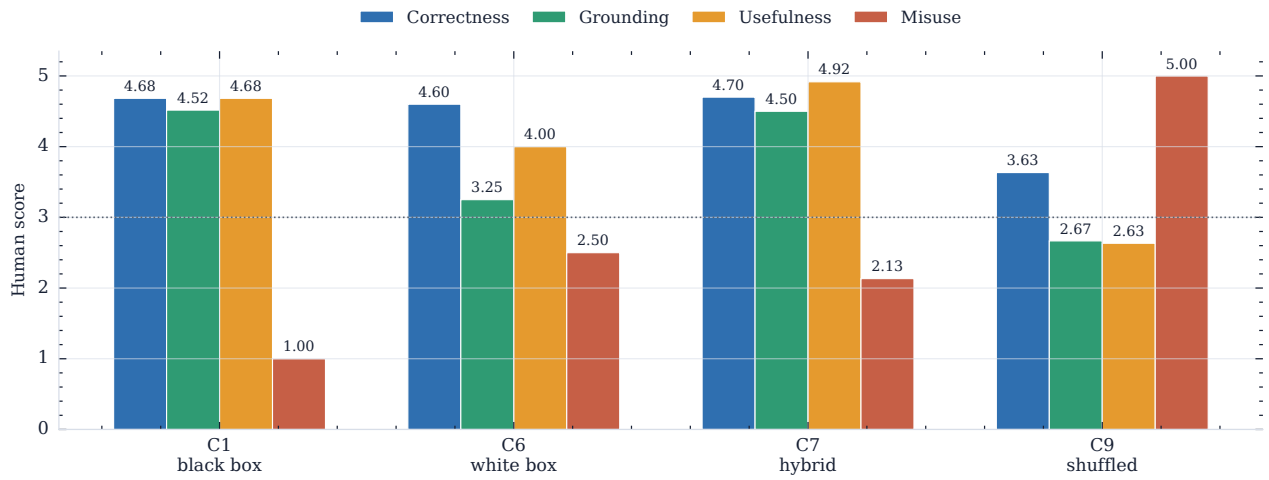


*Figure 2.* Primary human review scores for C1, C6, C7, and C9 over the 60 case evaluation set. Scores use a 1 to 5 scale; higher is better except evidence misuse.

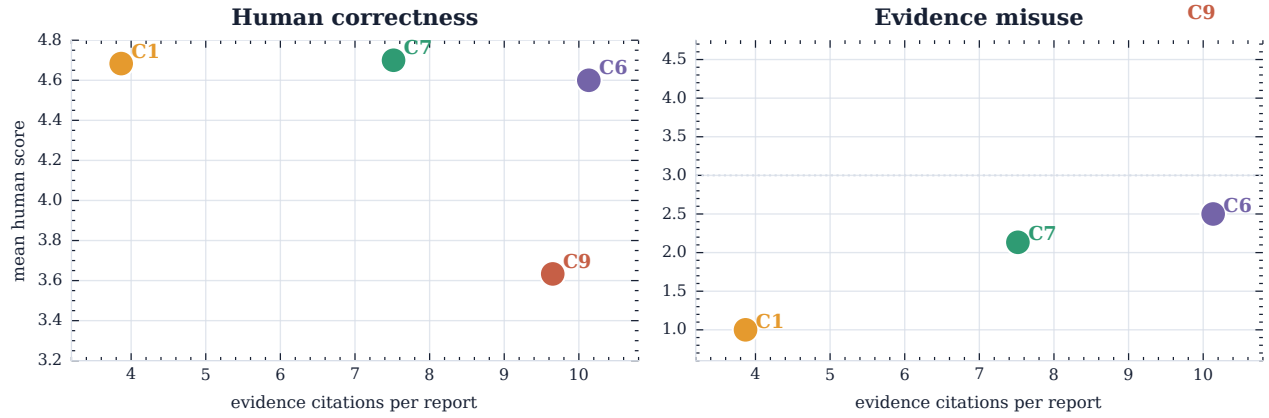


*Figure 3.* Citation volume plotted against human correctness and evidence misuse. The x axis reports mean evidence citations per report over the 60 case run; the y axes report mean human review scores over the same cases.

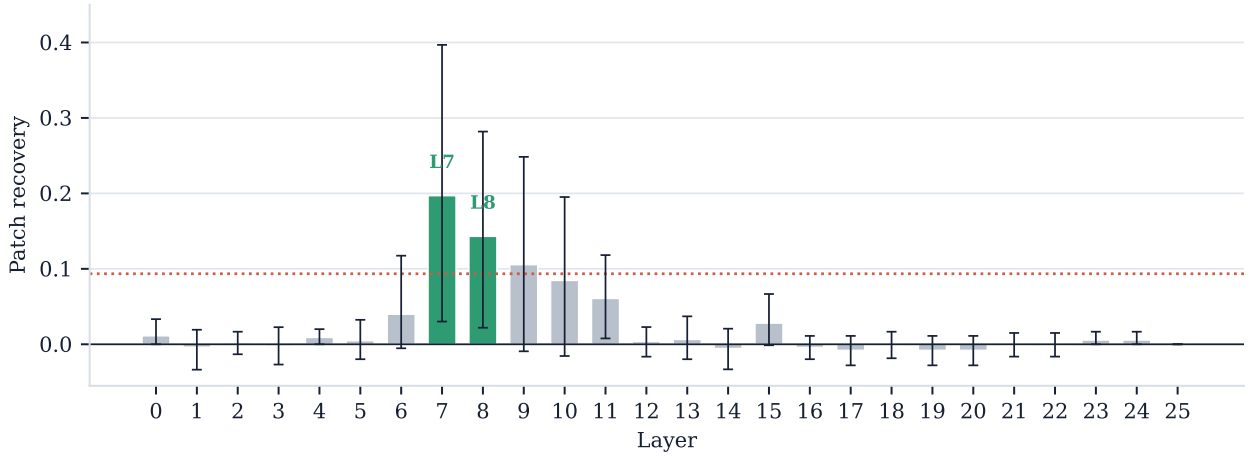


*Figure 4.* Residual stream activation patching diagnostic for deontic force. Bars show mean patch recovery by layer at the final prompt token; error bars show bootstrap intervals, and the dotted line marks the empirical causal gate.

## 7. Discussion and Limitations

The results support design lessons rather than a simple performance ranking. First, passage anchors should remain visible. Human review repeatedly identifies broad SAE labels, generic logit lens directions, and non specific steering evidence as failure modes when they are detached from the concrete actor, obligation, exception, or remedy in the passage. Second, citation volume should not be used as a quality measure. C6 and C9 both induce heavy citation behavior, but human review separates useful support from invalid support. Third, relevance controls are mandatory. If an LLM auditor can produce a plausible report under shuffled internal evidence, the interface risks transparency theater rather than reliable audit evidence.

The mechanistic interpretation is deliberately narrow. The auditor does not ignore the white box package: it cites SAE labels at very high per item rates and repeatedly reuses broad policy labels such as transparency, accountability, discrimination protection, and expert consultation. But the residual patching diagnostic shows why those citations should not be read as broad mechanistic support. The localized deontic force signal appears only at layers 7 and 8 at the final prompt token, whereas the report interface exposes many more readable labels and token directions. This pattern is consistent with a failure of evidence interpretation rather than a failure of evidence availability. The report treats internal records as topic cues, while human review asks whether those records actually support the specific passage claim.

The intended governance use is narrow. The paper does not propose an automated legal compliance system, and it does not test whether human auditors would be fooled by shuffled evidence. Instead, it evaluates evidence interfaces for LLM assisted audit workflows that are reviewed by humans or institutions: reviewers need to know which evidence was available, what was cited, and whether irrelevant evidence would have produced similar confidence. For technical AI governance, the contribution is an operationalization of access and verification. For mechanistic interpretability, the contribution is an application level evaluation of tool outputs through downstream report use and negative controls.

Several limitations point to the next experiments. The practical hypothesis is that anchoring matters, but the current design does not isolate anchoring from evidence quantity and variety. In the primary review records, package size is positively associated with evidence misuse (Spearman $\rho = 0.455$) but not with diagnostic usefulness ($\rho = -0.029$); this is not causal evidence, but it reinforces the need for volume controlled tests. A capped hybrid condition would separate anchoring from package size. A hybrid plus shuffled condition would test whether surface anchors protect against irrelevant internal evidence. Numeric only or label masked internal evidence would test whether readable labels drive over trust. A visible but not citable condition would separate evidence influence from citation behavior. Finally, running another auditor model or stricter prompt would test whether the pattern is Qwen specific.

The human review uses five reviewers and covers all 60 cases for the four primary conditions, but it should still be read as a diagnostic study rather than a final benchmark. Agreement diagnostics are measured on the original 20 case subset rather than all 60 cases, and single tool conditions C2 through C5 and raw AutoInterp C8 are human scored only on that subset. The target model is small, and the auditor is one local model. The activation explanation tool is a surrogate rather than a compatible real Activation Oracle. We do not calibrate SAE label quality, logit lens noise, or steering direction quality, so the experiment cannot fully separate interface risk from tool noise. The cleanup step fixes formatting and citation consistency only; it does not rewrite

substantive findings. Because reviewers score the cleaned reports, the live burden on an auditor may be higher. Finally, AGORA policy passages support policy audit research but do not by themselves define legal compliance judgments.

## 8. Conclusion

White box evidence packages change how an auditor LLM writes passage anchored policy audit reports, but full white box access alone does not improve grounded audit quality. It increases internal evidence use while weakening grounding and increasing misuse. The evidence is not ignored; it is often read as broad semantic support, even though the stricter patching diagnostic localizes causal support only narrowly. Hybrid packages are more promising in this design, but the result is confounded with package size and should not be read as proof that anchoring alone causes the gain. The shuffled control shows why this distinction matters: plausible reports can still misuse invalid internal evidence, and the report's own cautionary notes do not reliably catch that mismatch. The broader lesson is that internal access is valuable for policy auditing only when it is treated as evidence to be validated, not as authority to be trusted.

## Impact Statement

This work aims to improve transparency and evidence quality in policy relevant AI auditing. The main risk is over interpreting internal evidence as legal or causal proof. The paper therefore reports human review, negative controls, and claim boundaries explicitly.

## LLM Usage Statement

LLMs were used to assist with drafting, editing, formatting, code generation, debugging, summarizing reviewer feedback, and checking clarity. All experimental design decisions, data analysis decisions, human review, reported numbers, and final claims were reviewed and controlled by the author.

## References


Arnold, Z., Schiff, D., Schiff, K., Love, B., Melot, J., Singh, N., Jenkins, L., Lin, A., Pilz, K., Enweareazu, O., and Girard, T. Introducing the AI Governance and Regulatory Archive (AGORA): An analytic infrastructure for navigating the emerging AI governance landscape. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pp. 39–48, 2024. doi: 10.1609/aies.v7i1.31615.

Bell, A., Nov, O., and Stoyanovich, J. Think about the stakeholders first! towards an algorithmic transparency playbook for regulatory compliance. *arXiv preprint arXiv:2207.01482*, 2022. URL https://arxiv.org/abs/2207.01482.

Belrose, N., Furman, H., Smith, L., Halawi, D., Ostrovsky, I., McKinney, L., Biderman, S., and Steinhardt, J. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., Sutskever, I., Leike, J., Wu, J., and Saunders, W. Language models can explain neurons in language models. OpenAI, 2023.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023.

Burnat, F. A. D. and Davidson, B. I. The accountability paradox: How platform API restrictions undermine AI transparency mandates. *arXiv preprint arXiv:2505.11577*, 2025. URL https://arxiv.org/abs/2505.11577.

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., Sharkey, L., Krishna, S., Von Hagen, M., Alberti, S., Chan, A., Sun, Q., Gerovitch, M., Bau, D., Tegmark, M., Krueger, D., and Hadfield-Menell, D. Black-box access is insufficient for rigorous AI audits. In *ACM Conference on Fairness, Accountability, and Transparency*, 2024. doi: 10.1145/3630106.3659037. URL https://arxiv.org/abs/2401.14446.

Cen, S. H. and Alur, R. From transparency to accountability and back: A discussion of access and evidence in AI auditing. *arXiv preprint arXiv:2410.04772*, 2024. URL https://arxiv.org/abs/2410.04772.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024.

Cyberey, H., Ji, Y., and Evans, D. White-box sensitivity auditing with steering vectors. *arXiv preprint arXiv:2601.16398*, 2026. URL https://arxiv.org/abs/2601.16398.

Harack, B., Trager, R. F., Reuel, A., Manheim, D., Brundage, M., Aarne, O., Scher, A., Pan, Y., Xiao, J., Loke, K., Adan, S. N., Bas, G., Caputo, N. A., Morse, J. C., Ahuja, J., Duan, I., Egan, J., Bucknall, B., Rosen, B., Araujo, R., Boulanin, V., Lall, R., Barez, F., Alvira, S., Katzke, C., Atamli, A., and Awad, A. Verification for international AI governance. Oxford Martin AI Governance Initiative, 2025.

Holznagel, D. Shortcomings of the first DSA audits and how to do better. DSA Observatory, 2025.

International Auditing and Assurance Standards Board. International standard on assurance engagements 3000 (revised): Assurance engagements other than audits or reviews of historical financial information. International Federation of Accountants, 2013.

Karvonen, A., Chua, J., Dumas, C., Fraser-Taliente, K., Kantamneni, S., Minder, J., Ong, E., Sen Sharma, A., Wen, D., Evans, O., and Marks, S. Activation oracles: Training and evaluating LLMs as general-purpose activation explainers. arXiv preprint arXiv:2512.15674, 2025a.

Karvonen, A., Rager, C., Marks, S., et al. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *International Conference on Machine Learning*, 2025b.

Lieberum, T., Rajamanoharan, S., Conerly, A., Smith, L., Sonnerat, N., Varma, V., Kramar, J., Dragan, A., Shah, R., and Nanda, N. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *BlackboxNLP at EMNLP*, 2024.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Ojewale, V., Steed, R., Vecchione, B., Birhane, A., and Raji, I. D. Towards AI accountability infrastructure: Gaps and opportunities in AI audit tooling. *arXiv preprint arXiv:2402.17861*, 2025. URL https://arxiv.org/abs/2402.17861.

Reuel, A., Bucknall, B., Casper, S., Fist, T., Soder, L., Aarne, O., Hammond, L., Ibrahim, L., Chan, A., Wills, P., Anderljung, M., Garfinkel, B., Heim, L., Trask, A., Mukobi, G., Schaeffer, R., Baker, M., Hooker, S., Solaiman, I., Luccioni, A. S., Rajkumar, N., Moës, N., Ladish, J., Guha, N., Newman, J., Bengio, Y., South, T., Pentland, A., Koyejo, S., Kochenderfer, M. J., and Trager, R. Open problems in technical AI governance, 2024a. URL https://taig.stanford.edu.

Reuel, A., Soder, L., Bucknall, B., and Undheim, T. A. Position: Technical research and talent is needed for effective AI governance. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42543–42557. PMLR, 2024b. URL https://proceedings.mlr.press/v235/reuel24a.html.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., and Perez, E. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023. URL https://arxiv.org/abs/2310.13548.

Sheshadri, A., Ewart, A., Fronsdal, K., Gupta, I., Bowman, S. R., Price, S., Marks, S., and Wang, R. Auditbench: Evaluating alignment auditing techniques on models with hidden behaviors. *arXiv preprint arXiv:2602.22755*, 2026.

Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E. H., Scharli, N., and Zhou, D. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227. PMLR, 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

Solarova, S., Mesarcik, M., Pecher, B., and Srba, I. Beyond the checkbox: Strengthening DSA compliance through social media algorithmic auditing. *arXiv preprint arXiv:2601.18405*, 2026. URL https://arxiv.org/abs/2601.18405.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Transformer Circuits Thread, 2024.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, 2020.

Waiwitlikhit, S., Stoica, I., Sun, Y., Hashimoto, T., and Kang, D. Trustless audits without revealing data or models. *arXiv preprint arXiv:2404.04500*, 2024. URL https://arxiv.org/abs/2404.04500.

Yoran, O., Wolfson, T., Ram, O., and Berant, J. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*, 2024. URL https://arxiv.org/abs/2310.01558.

Zhang, F. and Nanda, N. Towards best practices of activation patching in language models. arXiv preprint, 2024.

## A. Full Evidence Condition Details

The experiment compares evidence interfaces, not retrieval systems. Each condition provides the same policy passage and audit rubric, but changes the auditor facing evidence items. Table 5 expands the condition table used in the main paper.

*Table 5.* Appendix condition details. Citation rules are part of the interface design: passage spans and evidence entries are separate objects.

| Condition | Auditor input | Allowed evidence citation | Expected misuse risk |
|---|---|---|---|
| C0 passage only | Policy passage, source information, and fixed rubric. | Passage spans only. | Claims without passage support may look plausible because no external evidence entries are available. |
| C1 black box surface | Passage plus surface observations, issue notes, selected spans, and confidence annotations. | Black box surface entries and passage spans. | Surface notes may encourage generic legal readings, but evidence remains tied to text. |
| C2 SAE only | Sparse feature activations, layers, activated spans, labels, review annotations, and caveats. | SAE entries and passage spans. | Readable feature labels may be treated as policy conclusions. |
| C3 logit lens only | Intermediate token direction records at selected layers and positions. | Logit lens entries and passage spans. | Token direction evidence may be overread as a causal explanation. |
| C4 steering only | Hidden state direction probes with alpha, token changes, and caveats. | Steering entries and passage spans. | Sensitivity evidence may be mistaken for a deployable control or legal proof. |
| C5 activation explanation surrogate | Natural language activation hypotheses derived from local evidence. | Surrogate explanation entries and passage spans. | Natural language hypotheses can sound authoritative despite being unvalidated. |
| C6 full white box | SAE, logit lens, steering, and activation explanation surrogate items. | All white box evidence entries and passage spans. | Dense internal evidence can distract from the passage and increase over citation. |
| C7 hybrid | C1 surface evidence plus C6 white box evidence. | Surface entries, white box entries, and passage spans. | Highest average usefulness in the primary review, but still vulnerable to internal evidence overuse. |
| C8 raw AutoInterp SAE | Raw SAE labels and feature references without the surrounding package context. | Raw SAE entries and passage spans. | Label readability may hide missing context and caveats. |
| C9 shuffled white box control | C6 style white box evidence from another case, hidden from the auditor. | Shuffled white box entries and passage spans. | Measures whether plausible internal evidence is trusted when relevance is broken. |

## B. Tool Construction Details

This appendix defines the evidence artifacts exposed to the auditor. These details matter because the experiment evaluates interfaces, not the intrinsic correctness of any one interpretability tool.

**Black box surface evidence.** C1 contains no hidden state information. The passage is split into sentences, and each sentence is scored by an audit cue count $b(s) = \sum_{\tau \in T} \mathbf{1}[\tau \text{ occurs in } s]$. The cue lexicon includes obligation, prohibition, risk, safety, security, transparency, monitoring, assessment, testing, threshold, compliance, enforcement, provider, operator, user, deploy, market, harm, vulnerable, and fundamental rights terms. The five highest scoring sentences become numbered surface evidence entries. Each item stores the sentence index, matched cue count, a supporting passage span, and the caveat that the item is extractive surface evidence.

**SAE evidence.** For Gemma 2 2B residual states, Gemma Scope SAE features are summarized as candidate concept activations. For feature $k$ at layer $l$, the package records

$$A_{i,l,k} = \max_p a_{i,l,k,p}, \qquad p^*_{i,l,k} = \arg\max_p a_{i,l,k,p}. \tag{1}$$

Features with nonzero activation are sorted by activation score, capped to a compact set, and rendered with layer, feature id, peak token, activated span, model id, SAE release, and caveat. In C2, C6, and C7, labels use human revised AutoInterp where the review marked the label usable or preserved. C8 exposes raw AutoInterp style labels without that review layer.

**Logit lens evidence.** Logit lens positions are selected from audit cue token offsets in the passage, with a fallback to the final active token. At layers $l \in \{6, 12, 18, 24\}$, the package computes $\ell_{i,l,p} = W_U \mathrm{LN}(h_{i,l,p})$ and records the top five vocabulary tokens and scores. These items are evidence about intermediate token directions, not evidence that the model finally predicts or endorses those tokens.

**Steering evidence.** Steering directions are constructed from positive prompts about obligations, risk controls, compliance duties, monitoring, transparency, accountability, enforcement, and harm protection, contrasted with negative prompts

describing background text without binding duties. For each layer $l \in \{12, 18, 24\}$,

$$v_l = \frac{\mu_l^+ - \mu_l^-}{\|\mu_l^+ - \mu_l^-\|_2}, \qquad h'_{i,l,p} = h_{i,l,p} + 0.35\|h_{i,l,p}\|_2 v_l. \tag{2}$$

The package records base and steered top tokens, whether the top token changes, and the KL divergence between base and steered vocab distributions. It does not run a full generation under intervention.

**Activation explanation surrogate.** The activation explanation condition is a conservative surrogate. It converts selected SAE labels, activations, peak tokens, and spans into a short activation meaning hypothesis with explicit caveats. The item is explicitly labeled as a surrogate so that the auditor cannot treat it as a validated Activation Oracle output.

**Shuffled control.** C9 keeps the target policy passage fixed but shifts full white box packages across cases. Thus $x_i$ is paired with $\mathcal{E}_{j,\mathrm{C6}}$ for $j \neq i$ while the auditor sees only a normal full white box shaped package. The control tests whether relevance is being checked or whether plausible internal evidence is being cited because it looks audit like.

## C. Audit Prompt and Report Schema

The auditor prompt is designed to make traceability a first class output rather than a later reconstruction. The prompt states that the policy passage is the primary source of truth and that evidence items are hypotheses or pointers. The auditor must return a structured report, cite exact short passage spans, avoid invented evidence citations, and avoid causal claims from one tool alone. Each report contains a concise summary; issue findings with actors, obligations or risks, support spans, cited evidence, confidence, and uncertainty; plausible alternative readings; internal evidence that materially influenced the report; cautionary notes about claims that should not be overstated; and an overall confidence score.

The schema separates three evidence objects: the policy passage, condition specific evidence entries, and the auditor's own claims. This separation is what makes shuffled evidence failures visible. A report can have correct passage claims while still citing invalid internal evidence.

## D. Human Gold Review Protocol

Human review has two stages, and five reviewers contribute across gold writing, verification, diagnostic scoring, and agreement checks. First, gold briefs are written from the policy passage and source information without seeing model reports or evidence packages. Each gold brief records issue tags, actors, obligations, support spans, known confounds, and claims to penalize when passage support is missing. Second, diagnostic review scores auditor reports against the gold brief for the four primary conditions: C1, C6, C7, and C9. The primary analysis covers all 60 cases. Agreement diagnostics use the original 20 case subset, and optional single tool review scores C2, C3, C4, C5, and C8 on that same subset.

For each reviewed report, the diagnostic vector is

$$h_{i,c} = (h_{i,c}^{\mathrm{corr}}, h_{i,c}^{\mathrm{ground}}, h_{i,c}^{\mathrm{use}}, h_{i,c}^{\mathrm{misuse}}). \tag{3}$$

The first three coordinates reward match to the gold brief, while the fourth penalizes invalid evidence use. This means a report can be high quality on passage interpretation and still fail as an audit evidence artifact if $h_{i,c}^{\mathrm{misuse}}$ is high. The C9 analysis focuses on exactly this conjunction.

Let $I$ be the reviewed case set and $K$ the four human review axes. The condition mean for axis $k$ is

$$\bar{h}_{c,k} = \frac{1}{|I|} \sum_{i \in I} h_{i,c,k}, \tag{4}$$

where $h_{i,c,k} \in \{1, \ldots, 5\}$. For pairwise statements relative to C1, we count better, tied, and worse outcomes on the same cases, reversing the interpretation only for misuse because lower misuse is better. We also report paired deltas $\Delta_{i,c,k} = h_{i,c,k} - h_{i,\mathrm{C1},k}$, exact sign tests, and Wilcoxon signed rank normal approximations as descriptive checks. For structural citation uptake, we report $\bar{U}_c = |\mathcal{D}|^{-1} \sum_i |Z_{i,c}|$. For the shuffled warning, we count cases satisfying $h_{i,\mathrm{C9,correct}} \geq 4$ and $h_{i,\mathrm{C9,misuse}} \geq 4$.

*Table 6.* Human review axes and claim boundaries.

| Axis | Measures | Does not measure |
|---|---|---|
| Correctness | Match to the gold issue, actor, and obligation structure. | Whether the target model actually used the cited internal feature causally. |
| Span grounding | Whether major claims are tied to concrete passage spans. | Whether the cited span is sufficient for a legal decision outside the passage. |
| Diagnostic usefulness | Whether the report would help an auditor notice the policy issue or gap. | Whether the report is a complete compliance analysis. |
| Evidence misuse | Whether the report overreads, overcites, or relies on irrelevant internal evidence. | Whether the passage level claim is necessarily wrong. |

*Table 7.* Agreement diagnostics on the original 20 cases and four primary conditions. QWK denotes quadratic weighted kappa.

| Axis | Reviewer 1 | Reviewer 2 | Exact | Within one | QWK |
|---|---|---|---|---|---|
| Correctness | 4.06 | 4.28 | 0.48 | 0.91 | 0.25 |
| Grounding | 3.25 | 3.71 | 0.48 | 0.91 | 0.50 |
| Usefulness | 4.08 | 4.00 | 0.53 | 0.93 | 0.46 |
| Misuse | 2.36 | 2.68 | 0.56 | 0.98 | 0.88 |

*Table 8.* Optional single tool human review on the original 20 cases. These results are secondary diagnostics and are not used as the main quality claim.

| Condition | Correct | Ground | Useful | Misuse |
|---|---|---|---|---|
| C2 SAE only | 3.45 | 3.15 | 3.00 | 3.25 |
| C3 logit lens only | 2.40 | 1.25 | 1.95 | 4.00 |
| C4 steering only | 3.35 | 3.25 | 3.00 | 3.00 |
| C5 activation surrogate | 3.30 | 3.00 | 3.00 | 2.70 |
| C8 raw AutoInterp SAE | 3.30 | 1.25 | 2.00 | 4.00 |

## E. All Condition Structural Summary

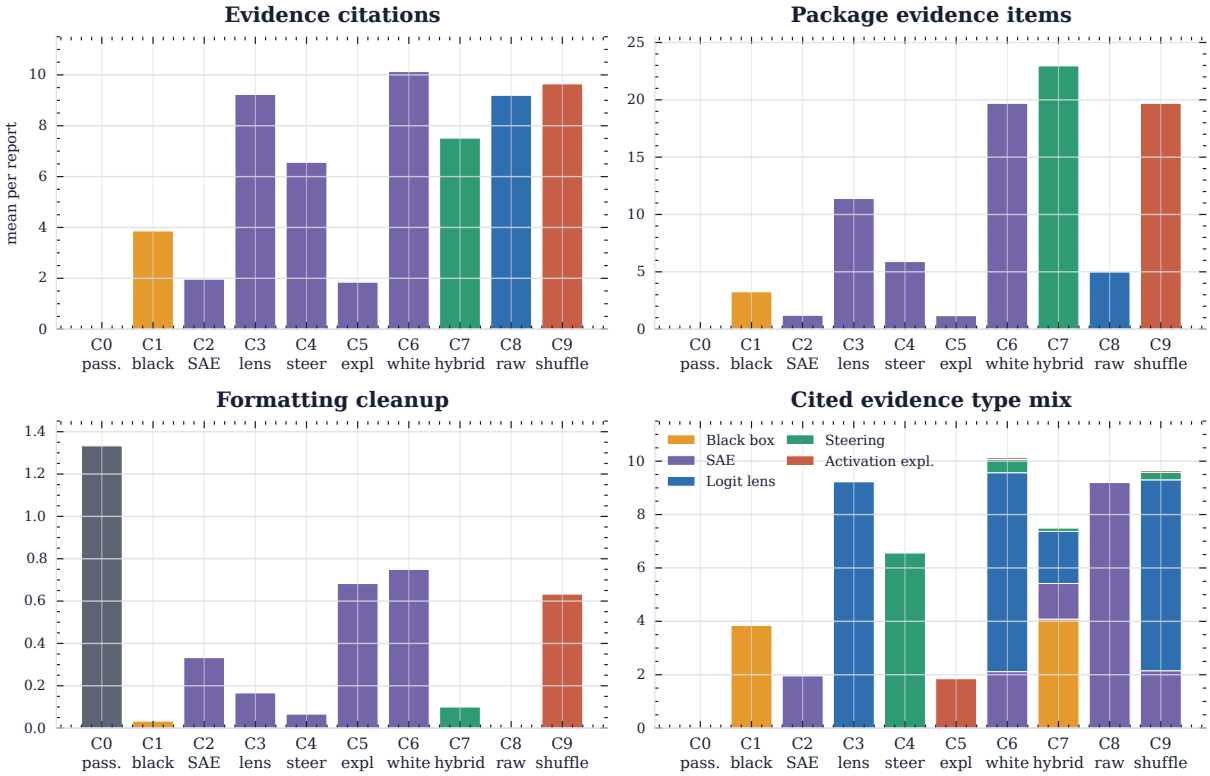


*Figure 5.* Structural metrics across all ten evidence conditions. Single tool conditions change citation behavior and formatting burden, but these measures are not human validity scores. C3 and C8 show high citation volume under single tool interfaces, while C9 shows that a full white box shaped package can be cited heavily even when relevance is broken.

## F. Case Study 1: Hybrid Anchoring Works

This case uses Articles 29 and 30 of the Internet Information Service Algorithmic Recommendation Management Provisions. The passage contains two distinct policy issues: confidentiality duties for bodies and personnel involved in supervision, and a complaint or report mechanism for organizations and individuals.

> Article 29: Related bodies and personnel participating in algorithmic recommendation service security assessment, supervision, and inspection shall maintain confidentiality of the personal private information, personal information, and commercial secrets they learn when exercising their duties and responsibilities, they may not disclose, sell, or illegally provide it to other persons.
> Article 30: Where any organization or individual discovers acts violating these Provisions, they may file a complaint or report with cybersecurity and informatization departments and relevant departments. Departments receiving complaints or reports shall handle them timely and according to the law.

The gold brief identifies confidentiality in supervision, complaint and report rights, and complaint handling obligations. Table 9 shows why C7 is a better interface for this case. The important point is not that white box evidence alone solves the interpretation. C6 narrows the report to confidentiality and legal liability, while C7 keeps the surface anchors that identify both Article 29 and Article 30.

*Table 9.* Case study 1 diagnostic comparison for a Chinese algorithmic recommendation provision on confidentiality and complaint handling. Scores are correctness, grounding, usefulness, and misuse. Lower misuse is better.

| Condition | Report behavior | Cited evidence pattern | Human interpretation |
|---|---|---|---|
| C1 | Identifies confidentiality and complaint handling, but grounding is not fully comprehensive. | Uses selected surface spans from the passage. | 5, 4, 5, 1. Strong passage anchored baseline. |
| C6 | Captures confidentiality but weakens grounding and usefulness relative to the surface baseline. | Uses logit lens and steering evidence around selected token positions. | 5, 3, 4, 2. Related internal evidence is not enough to preserve full passage structure. |
| C7 | Covers confidentiality, nondisclosure, complaint or report rights, and handling by departments. | Uses black box surface anchors while white box evidence remains secondary. | 5, 5, 5, 2. Highest scored result in this case; passage anchors preserve both Articles 29 and 30. |
| C9 | Produces a partially correct confidentiality finding but cites shuffled transparency evidence. | Uses shuffled SAE, logit lens, and steering evidence. | 4, 2, 3, 5. The internal evidence is mismatched even though the passage claim is partly correct. |

## G. Case Study 2: White Box Can Misframe Legal Category

This case uses an EU AI Act provision on high risk classification exceptions. The passage is legally delicate because it is not a generic operator duty. It specifies when an Annex III system is not considered high risk, while preserving an override for profiling of natural persons.

> By derogation from paragraph 2, an AI system referred to in Annex III shall not be considered to be high risk where it does not pose a significant risk of harm to the health, safety or fundamental rights of natural persons, including by not materially influencing the outcome of decision making. The first subparagraph shall apply where any listed condition is fulfilled, including a narrow procedural task, improving a previously completed human activity, detecting decision making patterns without replacing or influencing the prior human assessment without proper human review, or performing a preparatory task. Notwithstanding the first subparagraph, an AI system referred to in Annex III shall always be considered to be high risk where the AI system performs profiling of natural persons.

The gold brief identifies a classification exception, a significant risk threshold, decision making influence, Annex III scope, and a profiling override. Table 10 shows a limitation of the hybrid result: even C7 can shift to imprecise legal wording. This case keeps the paper's claim boundary intact. Hybrid is promising on average, not a guarantee of legal category precision.

*Table 10.* Case study 2 diagnostic comparison for an EU AI Act provision on high risk classification exceptions. Scores are correctness, grounding, usefulness, and misuse. Lower misuse is better.

| Condition | Report behavior | Cited evidence pattern | Human interpretation |
|---|---|---|---|
| C1 | Captures non high risk exceptions and the profiling carve out. | Uses surface evidence tied to passage spans. | 4, 4, 4, 1. Correct but could state more clearly that this is classification criteria. |
| C6 | Preserves the broad classification topic but can still blur the legal category. | Uses logit lens evidence around selected token positions. | 4, 3, 4, 2. Partially correct, but internal evidence is weak support for the legal distinction. |
| C7 | Captures conditions and profiling but uses imprecise low risk wording. | Mainly uses surface passage evidence despite receiving white box evidence. | 4, 4, 4, 2. Hybrid helps but the legal category is less precise than not high risk. |
| C9 | Infers a risk assessment gap and misses part of the mandatory profiling carve out. | Uses shuffled deployment evidence and unrelated logit lens evidence. | 3, 2, 2, 5. Control evidence is mismatched and the key exception structure is weakened. |

## H. Shuffled Control Failure Table

Table 11 lists representative C9 reports from the 39 cases with correctness at least 4 and evidence misuse at least 4. These are the most important governance failures because a reviewer who checks only passage level correctness could miss the invalid internal evidence use.

*Table 11.* Representative C9 cases with plausible correctness and high evidence misuse.

| Case | Source | Gold issue | Shuffled evidence cited | Reviewer finding |
|---|---|---|---|---|
| CN 12 | Chinese AR Provisions | User notice, transparency, service mechanism disclosure. | Logit lens. | Correctness 4 and misuse 5: passage overlap is plausible, but internal evidence is invalid support. |
| CN 13 | Chinese AR Provisions | Non targeted option, switch off right, tag controls, explanation for major rights impacts. | SAE and logit lens. | Correctness 4 and misuse 5: readable shuffled evidence makes the report look more grounded than it is. |
| CN 14 | Chinese AR Provisions | Minor protection, beneficial content, harmful content restriction, addiction prevention. | SAE and logit lens. | Correctness 4 and misuse 5: surface conclusion is close, while cited internal evidence comes from another case. |
| CN 21 | Chinese AR Provisions | Confidentiality, complaint or report right, complaint handling obligation. | SAE, logit lens, and steering. | Correctness 4 and misuse 5: the report recovers part of the passage but relies on shuffled transparency evidence. |
| CN 24 | Chinese AR Provisions | Filing fraud sanctions, filing cancellation, administrative penalties. | SAE and logit lens. | Correctness 4 and misuse 5: close to the passage, but control evidence should not be treated as support. |
| CN 7 | Chinese AR Provisions | Information security, unlawful content handling, records preservation. | SAE. | Correctness 4 and misuse 5: generic SAE labels appear persuasive despite being mismatched. |
| EU 182 | EU AI Act | Right to explanation and high risk AI deployer obligation. | SAE and logit lens. | Correctness 4 and misuse 5: primary obligation is plausible, but control evidence is mismatched. |
| EU 231 | EU AI Act | Technical documentation, validation, logs, cybersecurity measures. | Logit lens. | Correctness 4 and misuse 5: mostly correct passage summary with invalid internal citations. |
| EU 39 | EU AI Act | Human oversight, role allocation, and monitoring obligations. | Logit lens and steering. | Correctness 4 and misuse 5: the report sounds grounded while internal evidence relevance is broken. |
| GOV 1396-4 | Governance principles | Governance process, accountability, or safety framework requirements. | SAE and logit lens. | Correctness 4 and misuse 5: a broad governance theme survives, but the cited evidence package is from another case. |

## I. Human Review Score Distributions

Table 12 reports the score distributions behind the 60 case means. Each vector lists counts for scores 1 through 5. The distributions make the same claim boundary visible: C7 has many high usefulness scores, but it also shifts misuse upward; C9 remains often plausible on correctness while all shuffled reports score 5 on misuse.

*Table 12.* Expanded human review score distributions over 60 cases. Each vector is counts for scores 1/2/3/4/5.

| Condition | Correct | Ground | Useful | Misuse |
|---|---|---|---|---|
| C1 | 0/0/0/19/41 | 0/0/0/29/31 | 0/0/0/19/41 | 60/0/0/0/0 |
| C6 | 0/0/0/24/36 | 0/4/37/19/0 | 0/0/0/60/0 | 0/30/30/0/0 |
| C7 | 0/0/0/18/42 | 0/0/4/22/34 | 0/0/0/5/55 | 0/52/8/0/0 |
| C9 | 0/1/20/39/0 | 0/27/26/7/0 | 1/20/39/0/0 | 0/0/0/0/60 |

## J. Reproducibility Notes

The artifact repository for this paper is available at https://anonymous.4open.science/r/agora-mi-C473/README.md.

The active report file contains 600 JSONL records, one for each condition and case combination. Structural summaries are computed from the final quality check CSV. Human review summaries are computed from the expanded 60 case primary review CSV, the paired delta CSV, the inter annotator agreement summary, and the optional single tool review summary. The main paper reports only values that appear in those summary artifacts. The appendix case studies use the human gold brief, diagnostic review sheet, and final structured reports for the named cases.

Citation validity is a structural check:

$$V_{i,c} = \mathbf{1}\left[Z_{i,c} \subseteq \{id(e) : e \in \mathcal{E}_{i,c}\}\right]. \tag{5}$$

The reported value 1.0 means cited entries belong to the supplied package after formatting cleanup. It does not mean that the cited evidence is substantively relevant, which is why C9 can have perfect citation validity and high human scored misuse. Cleanup is restricted to parsing, schema normalization, confidence normalization, and citation formatting; it does not change the policy passage, evidence package, or substantive report findings.