AI Audits: Comparative Analysis, Costs, and Auditor Fit

I. Executive summary

This memo compares how the EU's General-Purpose AI Code of Practice (COP) and key U.S. state bills (Illinois HB 3506, Michigan HB 4668, New York's RAISE A-6453-B, and California SB 53) approach independent assurance for frontier models. It also explains why recurring third-party audits—not transparency alone—provide important ex-ante verifiability and trusted artifacts for regulators, insurers, customers, and boards. This memo discusses a harmonized "superset" program that uses the COP's Model Report (with independent external model evaluations and independent external security reviews) as the canonical dossier to satisfy IL/MI's auditor-independence, access, and publication requirements while accommodating NY/CA's transparency-first regimes. The memo also assesses the recent trend of removing audit mandates (NY/CA), details the stakes of shifting to after-the-fact enforcement, and considers who might be best suited to conduct auditing, ultimately recommending a hybrid audit team. The memo closes with risks and mitigations (public-records exposure, box-ticking, duplication, capacity/conflicts).

II. Third-Party Audits

Several AI policy analysts and standards bodies recommend recurring independent evaluations/audits as a first step toward accountability: NTIA's accountability work discusses when independent evaluations should be required; NIST's AI RMF emphasizes continuous, independent measurement; OECD guidance on AI accountability points the same way; and the UK AI Safety Institute has discussed the growing role of third-party evaluators for frontier systems.[1] Cross-industry experience (post-blackout electric reliability; post-Dieselgate auto emissions) likewise shows that moving from internal certification to independent, auditable checks improves real-world safety outcomes.[2]

---

[1] Nat'l Telecomms. & Info. Admin. (NTIA), *Independent Evaluations* (Mar. 27, 2024), https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report/developing-accountability-inputs-a-deeper-dive/ai-system-evaluations/independent-evaluations; Nat'l Inst. of Standards & Tech., *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* 26–34 (Jan. 2023), https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf; OECD, *Advancing Accountability in AI* (2023), https://oecd.ai/en/accountability; AI Safety Institute (UK), *Early Lessons from Evaluating Frontier AI Systems* (Oct. 24, 2024), https://www.aisi.gov.uk/work/early-lessons-from-evaluating-frontier-ai-systems.

[2] 2 U.S.–Canada Power Sys. Outage Task Force, *Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations* 1 (Dec. 2003), https://energy.gov/sites/prod/files/oeprod/DocumentsandMedia/BlackoutFinal-Web.pdf; Eur. Parl., Comm. of Inquiry, *Report on Emission Measurements in the Automotive Sector* ¶¶ 142–45 (2017), https://www.europarl.europa.eu/doceo/document/A-8-2017-0049_EN.html.

The core case for third-party compliance reviews in frontier AI is twofold: they increase adherence to safety frameworks and they produce assurance artifacts that external and internal stakeholders will actually trust. External audiences (regulators, downstream developers, customers, the broader public) are more likely to credit independent findings than self-assessments; public summaries can also reduce "race" dynamics by easing pressure to ship untested models. Internally, boards, governance bodies, and staff get a structured readout that demonstrates real gaps and drives resourcing and follow-through.

In recent years, several regulatory instruments have attempted to embed third-party auditing or independent external evaluation into AI governance. These include the EU General-Purpose AI Code of Practice, which steers providers to include outputs of independent external model evaluations and independent external security reviews in a single Model Report; Michigan H.B. 4668 (2025), which requires third-party audits with relevant technical expertise; and Illinois H.B. 3506 (2025), which mandates reputable third-party reviews with broad access and public reporting.[3]

III. EU AI Act

The EU AI Act was adopted on June 13, 2024 and published in the Official Journal on July 12, 2024 as Regulation (EU) 2024/1689.[4] For general-purpose AI (GPAI), the Commission published a voluntary Code of Practice on July 10, 2025. The Commission and the AI Board describe it as an adequate voluntary tool for showing you comply with Articles 53 (GPAI transparency/documentation) and 55 (extra duties for GPAI with systemic risk). GPAI obligations begin to apply from August 2, 2025 (with an additional runway for models already on the market to come into line by August 2, 2027).[5]

Major frontier providers have signed the Code of Practice. As of September 1, 2025, the Commission's signatory list includes Amazon, Anthropic, Cohere, Google, IBM, Microsoft, Mistral AI, OpenAI, ServiceNow, Aleph Alpha, and others; xAI has signed only the Safety & Security chapter (so it must demonstrate transparency/copyright compliance via other adequate means).[6]

It is important to distinguish what is legally binding in the Act from what the Code recommends. The only explicit third-party "audits" mandated by the Act sit in the high-risk

[3] European Commission, *The General-Purpose AI Code of Practice* (July 10, 2025), https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai; H.B. 4668, Reg. Sess. (Mich. 2025), https://legiscan.com/MI/bill/HB4668/2025; H.B. 3506, 104th Gen. Assemb., Reg. Sess. (Ill. 2025), https://legiscan.com/IL/bill/HB3506/2025.
[4] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on artificial intelligence, 2024 O.J. (L) 825.
[5] EU COP, supra note 3.
[6] Id.

regime: when an AI system is classified as high-risk, conformity assessment involves a quality-management system (QMS) review and periodic audits by a notified body under Annex VII. By contrast, GPAI obligations in Articles 53 and 55 do not themselves impose a one-size-fits-all external audit requirement.[7]

Where, then, does "auditing" show up for GPAI? In the Code of Practice—as credible ways to evidence your Article 53/55 compliance. The Code's Measure 7.4 asks providers to include in their Model Report any available outputs from independent external evaluators (Appendix 3.5) and independent external security reviews (Appendix 4.5), or to explain their absence. Appendix 3.5 sets expectations for independent third-party model evaluations—qualified evaluators, defined access/resources, and non-interference—to substantiate the systemic-risk evaluation and mitigation duties in Article 55. Appendix 4.5 recommends regular, independent external security reviews, complemented by red-teaming and bug bounties, to validate the effectiveness of security mitigations. None of these are stand-alone legal audit mandates; they are assurance practices that generate artifacts regulators can rely on.[8]

Why should frontier AI labs follow the Code at all if it is voluntary? Because the Commission explicitly frames adherence as a recognized pathway that offers more legal certainty than ad-hoc approaches when demonstrating compliance with Articles 53 and 55—even though adherence is not conclusive proof. In other words, the Code translates broad statutory duties into actionable, reviewable evidence that you can surface to the AI Office and national authorities.[9]

The AI Act's high-risk track hinges on deployment use-cases (see Annex III) and wraps them in product/QMS-style conformity assessment (including notified-body audits). By contrast, the state bills I'll analyze below (CA, NY, IL, MI) are developer-focused and emphasize organizational transparency, replicable test logs, and, in some states, annual third-party compliance audits. In this memo I focus solely on GPAI: the EU's binding "audit" concept lives in the high-risk/QMS world, while GPAI relies on documentation, evaluations, and security assurance—with the Code of Practice providing the clearest blueprint for external evaluations and security reviews that will help you satisfy scrutiny under Articles 53 and 55.[10]

IV. State Regimes

Four U.S. state proposals aimed at "frontier" AI developers frame the current landscape: California SB 53, Illinois HB 3506, Michigan HB 4668, and New York's RAISE Act (A6453). Illinois and Michigan are the audit-forward models: both would require at least annual,

---

[7] EU AI Act Explorer, *Annex VII: Conformity Based on Assessment of the Quality Management System and the Technical Documentation*, https://artificialintelligenceact.eu/annex/7/ (last visited Sept. 18, 2025).
[8] EU COP, supra note 3.
[9] Id.
[10] EU AI Act Explorer, *Annex III: Conformity Assessment Based on Self-Assessment (for Some AI Systems)*, https://artificialintelligenceact.eu/annex/3/ (last visited Sept. 18, 2025).

independent third-party audits beginning in 2026, broad auditor access to "all materials reasonably necessary," and publication of the full audit report within ninety days with narrow redactions and unredacted access for the Attorney General. By contrast, New York's current B version removed its earlier audit mandate and centers on publishing a Safety & Security Protocol, maintaining replicable testing logs, annual protocol reviews, and 72-hour incident disclosures; California likewise recently amended SB 53 to delete its delayed audit requirement. All four target large or frontier developers and vest enforcement with the state Attorney General, but they now diverge sharply on whether independent audits are required, what must be published, and who sees unredacted materials. In this section I focus on the audit-forward bills (Illinois and Michigan) and how a single program can satisfy both and align with the EU Code of Practice; in the next section I return to New York and California's move away from audits and what that means for compliance and oversight.

A. Illinois — HB 3506 (104th GA, 2025)

Illinois HB 3506 (104th GA, 2025) would require each covered developer to retain a reputable third-party auditor at least once every calendar year to assess whether the developer complied with its Safety & Security Protocol, flag any instances of noncompliance or ambiguous compliance, identify places where the protocol is not stated clearly enough to tell if the developer complied, and note any instances suggesting violations of the Act's truthfulness and redaction rules; developers must give auditors access to all materials produced to comply with the Act and any other materials reasonably necessary for the assessment, and they must conspicuously publish the full audit report within 90 days of completion, with any redactions justified in the published version while an unredacted copy is retained for five years and made available to the Attorney General; violations of the audit and publication provisions are enforceable by the Attorney General through injunctive relief and civil penalties up to $1,000,000. [11]

B. Michigan — HB 4668 (introduced Jun 24, 2025)

Michigan HB 4668 (introduced June 24, 2025) likewise mandates that, beginning January 1, 2026, large developers undergo not less than annual third-party audits in which the auditor evaluates compliance with the developer's Safety & Security Protocol, flags unclear protocol provisions that frustrate compliance determinations, and identifies possible violations of the Act's truthfulness and redaction provisions; developers must provide auditors access to all materials produced under the Act and any other materials reasonably necessary for the audit, must conspicuously publish the audit report within 90 days, and must ensure the auditor employs or contracts individuals with both corporate-compliance expertise and technical expertise in foundation-model safety; the Act also adopts a parallel redaction framework (including five-year retention of unredacted versions available to the Attorney General) and authorizes the Attorney

[11] Ill. H.B. 3506.

General to seek injunctive relief and civil fines up to $1,000,000 per violation of the audit and publication requirements.[12]

C. Harmonization

Labs push back on a state-by-state patchwork because it multiplies fixed compliance costs, creates conflicting obligations, and injects legal and pricing uncertainty into nationwide model releases. Different definitions of "frontier developer," divergent disclosure and publication rules, and non-uniform incident-reporting clocks force companies to build parallel processes and reformat the same evidence repeatedly; insurers and reinsurers then underwrite to the strictest regime, raising costs system-wide, while plaintiffs can forum-shop to states with broader disclosures or looser causation standards. These concerns show up repeatedly in federal analyses of AI policy, which warn that a patchwork of state laws can burden businesses and complicate implementation absent a coherent national framework.[13] The issue fed directly into this summer's "10-year moratorium" debate in Congress: some lawmakers proposed preempting state and local AI laws for a decade to avoid fragmentation, but the moratorium language was ultimately stripped from the larger package after intra-party and state-sovereignty pushback, leaving states free to proceed with their own rules.[14] Pro-moratorium commentators argued that without preemption the United States would drift toward inconsistent state regimes; opponents countered that a blanket pause would leave a regulatory vacuum.[15] Either way, the failure of preemption means firms now plan for—and price to—a patchwork.[16]

a. The "Superset" Compliance Program

One way to approach the labs' concerns about a difficult to manage state-by-state patchwork regime is to look at how difficult it really is to comply with the existing (or soon to be implemented) third-party auditing requirements. In this case, how might an AI lab satisfy the audit-forward state bills in Illinois and Michigan while also aligning with the EU GPAI Code of Practice?

Labs can do this by building a single dossier for each model family and each release, using the EU GPAI Code's Model Report as the template. The Model Report already expects a

[12] Mich. H.B. 4668.
[13] Cong. Rsch. Serv., R48555, Artificial Intelligence: Considerations for Addressing Risk and Regulation (2024), https://www.congress.gov/crs-product/R48555.
[14] Matt O'Brien, California Lawmakers Pass Landmark AI Safety Bill, Associated Press (Sept. 12, 2025), https://apnews.com/article/97d700da09cac62aa510eb4411bab24e.
[15] Steven Scheer, U.S. Senate Debates Whether to Adopt Revised State AI Regulation Ban, Reuters (June 30, 2025), https://www.reuters.com/business/media-telecom/us-senate-debates-whether-adopt-revised-state-ai-regulation-ban-2025-06-30/; Tambudzai Gundani., AI Regulation and Federalism: What the Moratorium (That Wasn't) Debate Revealed, GW Reg. Stud. Ctr. (Sept. 15, 2025), https://regulatorystudies.columbian.gwu.edu/ai-regulation-and-federalism-what-moratorium-wasnt-debate-revealed.
[16] Hodan Omaar, Without a Federal Moratorium, U.S. AI Policy Will Fragment Further, Ctr. for Data Innovation (July 15, 2025), https://datainnovation.org/2025/07/without-a-federal-moratorium-us-ai-policy-will-fragment-further/.

published Safety & Security Protocol and evidence from independent checks, so it provides a natural home for the core artifacts: the protocol itself, replicable testing logs, analysis of systemic risks and mitigations, and the outputs of independent model evaluations and independent security reviews.

On top of the EU-native dossier, labs can attach the state elements so the same package works in Illinois and Michigan. They can keep a rolling public update "no less than every 90 days" (IL) and "not less than once every 90 days" (MI); retain "a reputable third-party auditor" at least annually with "access to all materials … reasonably necessary" (IL/MI); and staff the audit team to "employ … individuals with expertise in corporate compliance" and "technical expertise in the safety of foundation models" (MI). They can then "conspicuously publish" the audit "no later than 90 days" after completion (IL/MI), "record and retain … for 5 years" the specific tests and results (IL/MI), and "allow the Attorney General to inspect" an unredacted version "upon request," while the published version "describe[s] the character and justification of the redaction" (IL/MI).[17]

With this workflow, labs publish once in the EU-style format, satisfy the annual audit and 90-day publication rules in Illinois and Michigan, and maintain unredacted annexes available for regulator inspection. In short, they operate with one dossier, one audit cycle, and one evidence trail—rather than parallel systems.

b. A Pragmatic Baseline

A second approach to addressing AI labs' concerns around compliance costs of a patchwork regime is to consider the minimum viable baseline regulation that states or Congress might adopt. A pragmatic path that many states could accept is a transparency-first baseline with targeted, risk-triggered assurance. The baseline would mirror the New York and amended California models: publish a Safety & Security Protocol with narrowly justified redactions, maintain test methods and results at a level that permits replication, conduct annual reviews, and report defined safety incidents within 72 hours while giving regulators confidential access to unredacted materials. On top of that, the statute would authorize regulators to require an independent external evaluation or limited-scope assurance engagement when clearly defined triggers are met—such as a major capability upgrade, a material change in model architecture or deployment scope, a prior safety incident, or evidence that logs are not in fact replicable. This preserves administrative simplicity for most developers, restores verifiability where stakes are high, and keeps states broadly aligned so providers can operationalize a single, durable compliance playbook. This baseline notably does not include a third-party auditing requirement. As I will discuss below, this is because there has been a trend of removing third-party auditing requirements in order to ensure smoother passage of the various AI regulation bills.

[17] Ill. H.B. 3506; Mich. H.B. 4668.

V. Trend of Auditing Requirement Removal

A notable shift this session is the quiet retreat from hard audit mandates to softer transparency regimes. California's SB 53 is the clearest example: earlier drafts required an annual, independent third-party audit beginning January 1, 2030, with specified auditor qualifications, "all materials reasonably necessary" access, five-year report retention, and a 30-day summary to the Attorney General. The September 2 amendment removed that requirement. What remains is a developer-authored Safety & Security Protocol plus record-keeping and disclosure duties. New York's RAISE Act followed a similar arc. The A6453-A draft included an annual third-party audit, an internal-controls review, and a signed lead-auditor certification; the current A6453-B drops the audit mandate and centers the regime on transparency: a published (redacted, with justification) Safety & Security Protocol, detailed testing logs sufficient for third-party replication, annual protocol reviews, and 72-hour safety-incident reporting to the Attorney General and DHSES.

Since the removal of the third-party audits, both bills rely on self-attestation backed by after-the-fact enforcement. That makes them easier to implement and less risky for trade secrets, but it also reduces ex-ante verifiability. In California, the amended bill no longer requires an independent party to attest that the Safety & Security Protocol is being followed in practice, which weakens deterrence against "safety-washing." New York's approach is similar: it creates real visibility via replicable test logs, publication, and rapid incident reporting, yet it stops short of requiring an external reviewer to validate control quality or test rigor. In both states, regulators retain the ability to scrutinize unredacted materials, but the removal of a recurring, independent check means the default assurance hinges on the developer's own documentation rather than on a standing, third-party examination.

What's at stake when recurring third-party audits disappear (or arrive only after crises)? While we can't be sure of how third-party audits might strengthen the safety of the AI sector, we can look at other industries and how they were impacted by auditing requirements. When sectors removed independent checks or leaned too hard on self-attestation, the pattern has been predictable: (i) errors go undetected until after harm; (ii) assurance artifacts lose comparability (insurers price for opacity); and (iii) regulators have to reconstruct the truth post hoc with weak evidence trails. The following are a few case studies of the impact of third-party auditing requirements.

- Electric grid (2003 Northeast blackout): Before Congress empowered FERC to make NERC reliability standards mandatory (and enforceable with penalties), the U.S.–Canada task force concluded "compliance with reliability rules must be made mandatory," precisely because voluntary, self-policed rules failed to prevent a continent-scale

outage.[18] The immediate fix was a regime with enforceable standards and audits/monitoring, a turning point Congress' later reviews underline.[19]

- Aviation certification (Boeing 737 MAX): Heavy delegation (FAA's ODA) without sufficiently independent oversight let safety-critical assumptions go unchallenged. DOT's Inspector General and GAO both found FAA processes and oversight of Boeing's delegated work ineffective, spurring reforms to tighten independent review.[20] Recent FAA enforcement since the Alaska MAX-9 door-plug failure underscores how costly weak, non-independent quality systems become after incidents.[21]
- Auto emissions (Dieselgate): The EU's pre-2015 type-approval system relied on lab tests run by services paid by manufacturers, with virtually no independent on-road checks—a conflict EU inquiries later flagged. Post-scandal, Europe added independent market surveillance and real-driving emissions verification to restore trust.[22]
- Public company accounting (pre-SOX): Self-regulation of auditors collapsed with Enron/WorldCom. Congress created the Public Company Accounting Oversight Board (PCAOB) to oversee external auditors and required auditor attestation of internal controls (SOX §404), precisely to replace self-attestation with independent assurance.[23]
- Crypto (FTX): The CEO overseeing the bankruptcy estate reported a "complete failure of corporate controls"; many financial statements were unaudited or of dubious

---

[18] U.S.–Canada Power Sys. Outage Task Force, *Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations* 5 (Apr. 2004), https://www.ferc.gov/sites/default/files/2020-05/ch1-3.pdf

[19] U.S. Gov't Accountability Off., GAO-04-204, *Electricity Restructuring: 2003 Blackout Identifies Crisis and Opportunity for the Electricity Sector* 5–6 (Nov. 2003), https://www.gao.gov/products/gao-04-204

[20] U.S. Dep't of Transp., Off. Inspector Gen., AV2021020, *Weaknesses in FAA's Certification and Delegation Processes Hindered Its Oversight of the 737 MAX 8* 1–3 (Feb. 23, 2021), https://www.oig.dot.gov/sites/default/files/FAA%20Certification%20of%20737%20MAX%20Boeing%20II%20Final%20Report%5E2-23-2021.pdf; U.S. Gov't Accountability Off., GAO-22-104480, *Aircraft Certification: FAA's and EASA's Processes* 1–3 (June 30, 2022), https://www.gao.gov/assets/gao-22-104480.pdf.

[21] David Shepardson, *FAA proposes to fine Boeing $3.1 million over widespread safety violations*, Reuters (Sept. 12, 2025), https://www.reuters.com/world/faa-proposes-fine-boeing-31-million-over-widespread-safety-violations-2025-09-12/

[22] Eur. Parl., Comm. of Inquiry, *Report on Emission Measurements in the Automotive Sector* ¶¶ 142–45 (2017), https://www.europarl.europa.eu/doceo/document/A-8-2017-0049_EN.html; Transport & Env't, *Dieselgate: Who? What? How?* 4–6 (Sept. 12, 2016), https://www.transportenvironment.org/uploads/files/2016_09_Dieselgate_report_who_what_how_FINAL_0.pdf; Eur. Ct. of Auditors, *The EU's Response to the "Dieselgate" Scandal* 4–6 (2019), https://www.eca.europa.eu/lists/ecadocuments/brp_vehicle_emissions/brp_vehicle_emissions_en.pdf; BEUC, *Volkswagen Emission Affairs* (2018), https://www.beuc.eu/volkswagen-emission-affairs.

[23] See, e.g., Daniel L. Goelzer, *Lessons from Enron: The Importance of Proper Accounting Oversight*, PCAOB (May 12, 2006), https://pcaobus.org/news-events/speeches/speech-detail/lessons-from-enron-the-importance-of-proper-accounting-oversight_45; William J. McDonough, *Sarbanes-Oxley and the Post-Enron Environment: Auditor Oversight*, PCAOB (Nov. 19, 2004), https://pcaobus.org/news-events/speeches/speech-detail/sarbanes-oxley-and-the-post-enron-environment-auditor-oversight_120; Sec. & Exch. Comm'n, *Study of the Sarbanes–Oxley Act of 2002 Section 404 Internal Control Over Financial Reporting Requirements* 16–20 (2009), https://www.sec.gov/news/studies/2009/sox-404_study.pdf

quality—contributing to billions in losses and chaotic recovery. The SEC later charged the auditor with misconduct.[24]

- Large-scale software updates (CrowdStrike, 2024): While not a "third-party audit" case, the global outage illustrates the system-wide cost of inadequate independent pre-release checks: experts reported skipped QA gates; insured losses alone were estimated at $400M–$1.5B.[25]

Removing recurring audits shifts risk onto the public and places an additional burden onto regulators, who must reconstruct compliance without independent artifacts. That is why many technical frameworks (e.g., NIST AI RMF) and the EU GPAI Code of Practice emphasize independent external evaluations/reviews as evidence you can rely on before harm.[26]

VI. Who Should Audit?

Beyond understanding the auditing requirements themselves and their implications, another important element is considering who should conduct the auditing itself. A paper by Homewood et al., 2025 considered third-party compliance reviews and offered a practical framework for deciding who should conduct them. The paper makes two core claims: (i) independent reviews can both raise adherence to a lab's own safety framework and create assurance artifacts that internal and external stakeholders trust; and (ii) reviewer choice is a trade-off among AI expertise, risk-management and compliance-review experience, public credibility, and the ability to handle sensitive information.[27] It also foregrounds the main failure modes (security leakage, cost, and false assurance) and treats internal and external reviews as complements rather than substitutes.[28] Finally, it organizes the space around six design questions, the first of which is precisely "who could conduct the review," and compares options such as the

---

[24] Tom Hals, *Bankrupt FTX's new CEO outlines fund abuses, slams "complete failure of corporate control"*, Reuters (Nov. 17, 2022), https://www.reuters.com/technology/new-ftx-ceo-slams-complete-failure-corporate-control-2022-11-17/; *SEC says FTX auditor did not understand the crypto market*, Fin. Times (2024), https://www.ft.com/content/ad04330b-4252-45d3-a7f9-b60a9d930081

[25] Zeba Siddiqui, *CrowdStrike update that caused global outage likely skipped checks, experts say*, Reuters (July 22, 2024), https://www.reuters.com/technology/cybersecurity/crowdstrike-update-that-caused-global-outage-likely-skipped-checks-experts-say-2024-07-20/; *Insured losses from CrowdStrike outage could reach $1.5 bln*, Reuters (July 25, 2024), https://www.reuters.com/business/finance/insured-losses-crowdstrike-outage-could-reach-15-bln-cybercube-says-2024-07-25/

[26] Nat'l Inst. of Standards & Tech., *AI Risk Management Framework 1.0* 26–34 (Jan. 2023), https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf; Eur. Comm'n, *The General-Purpose AI Code of Practice* (July 10, 2025), https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai; *Code of Practice—Appendix 3.5 & 4.5 overview*, EU AI Act Explorer (July 2025), https://artificialintelligenceact.eu/code-of-practice-overview/.

[27] Aidan Homewood et al., *Third-Party Compliance Reviews for Frontier AI Safety Frameworks* 2 (May 3, 2025), https://arxiv.org/abs/2505.01643 (describing benefits—raising adherence and producing trusted assurance artifacts—and principal challenges/failure modes).

[28] Id. at 8–11.

Big Four accounting firms, AI evaluation organizations, security firms, consulting and law firms, with "minimalist → comprehensive" pathways as capacity grows.[29]

Translating that framework to the audit-forward state bills (Illinois HB 3506 and Michigan HB 4668), the most defensible pattern is a hybrid team: a Big Four–style assurance firm as the lead auditor, paired with a frontier-model evaluation specialist as a named technical subcontractor, and—where scope includes infrastructure hardening and incident-response controls—a security audit firm for operational depth. The Big Four option scores highest on risk-management, compliance-review experience, and stakeholder credibility, but is thinner on frontier-AI specifics; the evaluator brings the needed model-safety expertise; the security firm contributes mature practices for handling sensitive artifacts.[30] This division of labor mirrors the paper's comparative matrix across reviewer types (AI expertise; risk-management expertise; compliance-review expertise; public reputation; ability to handle sensitive information) and avoids asking any single provider to over-extend beyond its core competency.[31]

This hybrid is also the cleanest fit to statutory text. Illinois HB 3506 calls for a "reputable" third-party auditor with broad access and publication of the full report within 90 days; a Big Four–led engagement plainly satisfies "reputable," while a named evaluator ensures the report can speak credibly to foundation-model safety claims the auditor is assessing.[32] Michigan HB 4668 goes further: auditors must employ or contract individuals with both corporate-compliance expertise and technical expertise in the safety of foundation models; developers must maintain replicable test logs; the public report is due in 90 days; and the Attorney General must be able to inspect unredacted materials, with the public version carrying redaction justifications.[33] The hybrid team maps one-to-one onto these requirements: make the assurance firm the signing lead, and designate technical personnel from the evaluator to satisfy Michigan's expertise clause. This structure satisfies MI's subject-matter expertise clause and IL's reputability/access/publishing requirements while creating COP-compatible artifacts that can be reused in EU filings.

For New York (RAISE A6453-B) and California (SB 53), where recurring third-party audits are no longer mandated, the same reviewer structure can be used voluntarily to support transparency regimes (e.g., published Safety & Security Protocols and replicable testing logs). In those cases, scope the engagement to produce a concise public summary and a regulator-ready pack; have the lead partner sign the summary, with technical appendices authored by the evaluator. That approach preserves credibility without over-engineering the review relative to the

[29] Id. at 19.
[30] Id. at 12.
[31] Id.
[32] Ill. H.B. 3506.
[33] Mich. H.B. 4668.

lighter statutory obligations—and it aligns with the paper's advice to stage ambition and to treat internal and third-party checks as complements.

VII. Risks, drawbacks, and mitigations

While third-party audits and transparency can raise trust and discipline, there are still important downsides. Public reporting can expose trade secrets or security details while making it more difficult for the public to understand risks; metrics can be gamed; overlapping EU–state rules can create duplicative work; and reviewer capacity/qualification is still not robust. This section lists those risks and pairs each with concrete mitigations so policymakers can have the benefits of assurance without creating new failure modes.

A. Confidentiality and Public Records Exposure

Publishing assurance artifacts and furnishing unredacted materials to agencies creates trade-secret and security exposure because public-records laws begin from a disclosure presumption: any record an agency possesses and controls is presumptively releasable.[34] Despite this presumption, there are some meaningful carve-outs. At the federal level, FOIA Exemption 4 protects "trade secrets and commercial or financial information obtained from a person [that is] privileged or confidential," and the Supreme Court has clarified that "confidential" bears its ordinary meaning—information customarily kept private and, often, provided under assurances of privacy—replacing the older "substantial competitive harm" gloss.[35] States provide analogous protections (e.g., California's CPRA exempts trade secrets and other specified records; New York's FOIL authorizes withholding where disclosure would cause "substantial injury to the competitive position"; Illinois and Michigan have similar provisions), but agencies must still release all reasonably segregable non-exempt material and, under the FOIA Improvement Act, consider foreseeable harm before withholding.[36]

Practically, a way to mitigate this might involve a two-track disclosure regime. Regulators can require AI labs to publish a redacted, plain-English summary of auditing results with line-item justifications while also placing unredacted annexes in regulator-only data rooms with clear Exemption-4 (and state-law) markings, consistent handling, and least-privilege access logs (using submitter-notice procedures and recognizing the limits of "reverse-FOIA" relief).[37] To further reduce leakage risk, require third-party teams to operate under standard confidentiality and independence duties drawn from the financial-audit world.[38]

---

[34] 5 U.S.C. § 552(a).
[35] 5 U.S.C. § 552(b)(4); *Food Mktg. Inst. v. Argus Leader Media*, 139 S. Ct. 2356 (2019).
[36] Cal. Gov't Code § 6254 (West 2021); N.Y. Pub. Off. Law § 87(2)(d) (McKinney 2024); 5 U.S.C. § 552(b); 5 U.S.C. § 552(a)(8)(A) (FOIA Improvement Act foreseeable-harm standard).
[37] Exec. Order No. 12,600, 3 C.F.R. 235 (1987); *Chrysler Corp. v. Brown*, 441 U.S. 281 (1979).
[38] AICPA, *Code of Professional Conduct* ET §§ 1.200, 1.700 (2025), https://pub.aicpa.org/codeofconduct/ethicsresources/et-cod.pdf; PCAOB Rule 3520, *Auditor Independence*, https://pcaobus.org/about/rules-rulemaking/rules/section_3.

B. Trade-Secret Redactions and Hollow Transparency

While protecting trade secrets can be important, another risk of the disclosure and auditing requirements is that it may obscure an SSP to the point that the public is unable to understand or scrutinize a model's risk profile. Even with Michigan's and Illinois's annual third-party audit mandates and regulator-only data rooms, auditors can end up reviewing substance that the public (and independent researchers) cannot evaluate, weakening ex-ante verifiability and shifting enforcement toward after-the-fact.[39]

In his LawFare article, Julius Hattingh suggests that a stronger mitigation is to pair the IL/MI audit mandates and regulator-only data rooms with two safeguards the article urges—(i) a public-interest override barring trade-secret redactions where disclosure is necessary to reasonably inform the public about catastrophic-risk controls, and (ii) a formal redaction-challenge workflow in which unredacted materials and reason-giving are submitted to the attorney general for review, with authority to order alternative disclosures (e.g., sanitized descriptions or aggregates).[40] Hattingh also notes a forward-looking risk that overprotecting secrecy today can entrench "reasonable investment-backed expectations" and complicate future efforts to tighten disclosure (Takings Clause), which reinforces adopting these guardrails now alongside least-privilege access, submitter-notice, and standard audit confidentiality to keep audits substantive without unnecessary leakage.[41]

C. Metric Gaming and "Box-Ticking"

A recurrent failure mode of auditing regimes is Goodharting: once the measure becomes the target, process can drift toward paperwork rather than risk reduction.[42] A way to keep audits from devolving into checklists is to anchor them to outcome-oriented indicators that track whether controls work in production (e.g., incident rates; mean time to detect and repair; rollback success rates) and to require periodic, independent re-execution of a sample of tests so results cannot overfit a static benchmark. This approach aligns with the "Measure/Manage" thrust of NIST's AI Risk Management Framework—which emphasizes continuous measurement, independent assessment, and feedback into governance—while keeping flexibility to adjust control sets and reviewers so they cannot be gamed.[43]

D. Cross-Jurisdiction Duplication and Evidence Fragmentation

---

[39] Julius Hattingh, New AI Transparency Rules Have a Trade Secrets Problem, Lawfare (Sept. 15, 2025, 3:21 PM), https://www.lawfaremedia.org/article/new-ai-transparency-rules-have-a-trade-secrets-problem.
[40] Id.
[41] Id.
[42] CNA, *Goodhart's Law*, CNA Analyses (Sept. 2022), https://www.cna.org/analyses/2022/09/goodharts-law.
[43] Nat'l Inst. of Standards & Tech., *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* 26–34 (Jan. 2023), https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

Without a common evidence base, compliance with EU expectations, state transparency duties, and any state audit mandates can turn into duplicative processes. A practical mitigation is to designate a single canonical dossier—using the EU General-Purpose AI Code of Practice ("COP") Model Report—and reuse it across jurisdictions (as discussed previously in the harmonization section). The COP orients providers to assemble independent external model-evaluation outputs and independent external security-review outputs (its Appendices 3.5 and 4.5), which map cleanly to what audit-forward U.S. state bills ask agencies to see. To minimize timing friction, run one annual cycle matched to the earliest state deadline, publish a harmonized public summary, and keep regulator-only annexes to satisfy broader access and publication rules in Illinois and Michigan.[44]

E. **Capacity, Independence, and Conflicts**

There is a scarcity of evaluators who combine enterprise-assurance practice with frontier-model safety and security depth. A hybrid team (as discussed previously) which could include a lead assurance firm subject to American Institute of Certified Public Accountants (AICPA) independence, workpaper retention, and conflicts rules, paired with a named frontier evaluation/security subcontractor that brings model-specific competence would likely be the best solution. This structure tracks general assurance norms and satisfies the direction of audit-forward state bills (Michigan's emphasis on subject-matter expertise; Illinois's reputability, broad-access, and publication expectations), while producing artifacts that can be dropped into the COP Model Report without duplication.[45] Finally, two points on independence and information handling. First, keep clear independence guardrails: do not hire the same evaluator that designed risk-critical pre-release tests to later grade them; avoid contingent/implementation fees; and apply a cooling-off period before hiring audit-team members. Second, implement least-privilege access and on-premises review rooms for the most sensitive items (e.g., weight-adjacent artifacts), with role segregation and audit-trail logging; these are standard mitigations from the paper to reduce security leakage and the risk of overly rosy or incorrect findings. Both would strengthen the credibility of an IL/MI-grade audit and are in line with the review's cautions and mitigations.

---

[44] European Commission, *The General-Purpose AI Code of Practice* (July 10, 2025), https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai; EU AI Act Explorer, *Code of Practice overview* (App. 3.5 & 4.5), https://artificialintelligenceact.eu/code-of-practice-overview/; H.B. 3506, 104th Gen. Assemb., Reg. Sess. (Ill. 2025), https://legiscan.com/IL/bill/HB3506/2025; H.B. 4668, Reg. Sess. (Mich. 2025), https://legiscan.com/MI/bill/HB4668/2025.

[45] AICPA ET §§ 1.200, 1.700; PCAOB Rule 3520; H.B. 4668 (Mich. 2025); H.B. 3506 (Ill. 2025), supra note 8.