A Blueprint for a Secure EU AI Audit Ecosystem

Alejandro Tlaie Boria [1 2]

Abstract

This paper proposes a blueprint for the European Union to transform its regulatory leadership into industrial advantage through a Secure EU AI Audit Ecosystem. While third-party AI audits are widely valued, we address the lack of a practical roadmap for building infrastructure that enhances the EU's competitiveness, digital sovereignty, and defence priorities. Our framework outlines a phased approach—from voluntary pilots to mandatory audits once feasibility is demonstrated—enabling accredited third parties to conduct secure, confidential evaluations of General-Purpose AI with systemic risk. Core components include legal IP protections, secure-computing technologies (whose current limitations we critically assess), and an oversight function for certifying auditors. Finally, we provide actionable implementation scenarios aligning with Chips Act facilities, AI (Giga)Factory clusters, and EuroHPC resources, estimating costs, outlining existing EU funding instruments, and proposing complementary policy measures to operationalize these proposals.

1. Motivation and Policy Gap

Artificial Intelligence has entered a new era characterized by increasingly capable general-purpose models such as the GPT series,the Gemini family, and Claude (Radford et al., 2019; Achiam et al., 2023; Hurst et al., 2024; Team et al., 2023; Georgiev et al., 2024; Anthropic, 2024; 2025). These systems present unprecedented complexity and opacity, raising serious societal concerns about bias, transparency, accountability, and potential misuse (Ferrer et al., 2021; Von Eschenbach, 2021). While regulatory efforts like the EU AI Act (Regulation - EU - 2024/1689 - EN - EUR-Lex) explicitly mandate model transparency, they fall short in providing clear guidelines for deeper, technical audits. Proprietary constraints, technological complexities, and geopolitical tensions further complicate effective oversight.

The EU's regulatory strategy has traditionally emphasized trust and human rights, but emerging General-Purpose AI with Systemic Risk (GPAISR), with wide-ranging and evolving capabilities, present challenges that exceed current legislative tools (Stix, 2022). The AI Act's risk-based approach does not sufficiently capture the full risk surface of general-purpose models whose applications cannot be exhaustively defined ex ante (Act, 2024). Similarly, the GDPR, while foundational for data governance, lacks operational mechanisms for auditing non-personal, system-level risks posed by AI models.

Despite commendable regulatory efforts, the EU currently lacks a secure, standardized framework for conducting the technical evaluations of complex AI models that would require deeper-than-black-box access (Casper et al., 2024). This gap allows potentially unsafe AI systems to operate without sufficient external oversight, undermining public trust and weakening EU competitiveness on the global stage. Against this backdrop, we propose a comprehensive framework that addresses this critical gap by enabling secure technical audits while balancing intellectual property concerns with public accountability needs.

Recent drafts of the *AI-Act Codes of Practice* (June 2025) already sketch voluntary transparency templates, yet stop short of prescribing the deep technical modalities required for GPAISR oversight. In parallel, the *Digital Services Act* Art. 37 mandates annual third-party audits for Very Large Online Platforms and Search Engines; if systems such as *ChatGPT* were designated, a GPAISR audit framework would supply the missing technical layer. Our proposal therefore bridges the gap between these horizontal obligations and the vertical needs of systemic-risk models.

2. Related Work

Previous work on AI auditing has largely focused on policy frameworks (Stix, 2022), verification processes (Shi et al., 2022), bias evaluation methodologies (Ferrer et al., 2021), and transparency mechanisms (Novelli et al., 2024). However, the technical infrastructure required to implement deep auditing of frontier AI models with adequate intellectual property protection remains underexplored.

[1]Talos Fellow [2]SaferAI. Correspondence to: Alejandro Tlaie Boria <atboria@gmail.com>.

Workshop on Technical AI Governance (TAIG) at ICML 2025, Vancouver, Canada.

Secure multi-party computation and confidential computing approaches have been proposed for protecting sensitive ML model information (Tramèr & Boneh, 2018; Tramèr et al., 2022; Islam, 2024; Gamiz et al., 2025), but these have not been operationalized for regulatory AI auditing. Similarly, others have noted the limitations of black-box access for rigorous AI evaluations (Casper et al., 2024), but practical solutions for deeper access while protecting IP remain limited. However, this is already changing, as OpenMined recently published an applied walk-through of enclave-based LLM evaluation (OpenMined, 2025), indicating growing practitioner interest.

Technical approaches like Trusted Execution Environments (TEEs) (Islam, 2024) and cryptographic solutions such as homomorphic encryption (Tramèr et al., 2022) offer potential pathways, but these have not been systematically applied to the AI auditing problem, only to simple cases (Tramèr & Boneh, 2018; Shi et al., 2022). Recent work on "trustless" audits demonstrates progress in secure evaluation without revealing sensitive data or model information (Waiwitlikhit et al., 2024; Tlaie & Farrell, 2025), though these approaches still face practical limitations in real-world regulatory contexts.

Audit disciplines in adjacent sectors offer instructive precedents: the European Medicines Agency conducts *Good Manufacturing Practice* audits of pharmaceutical plants; the Basel Committee coordinates bank *stress tests*; and (Clark & Hadfield, 2025) advocate competition among licensed auditors. Our framework incorporates their lessons on adversarial compliance incentives and fee-funded oversight.

3. Secure AI Audit Ecosystem Framework

We propose a three-pillar framework enabling accredited third-party auditors to perform deep, secure evaluations of advanced AI systems while protecting intellectual property.

3.1. Technical Architecture

The framework's core technical foundation leverages confidential computing technologies to create secure evaluation environments. The audit would follow a specific workflow (Fig. 3.1): **I) Enclave Verification:** Model providers first verify the integrity of the secure computing environment using cryptographic attestation (Niemi et al., 2022) before granting any access to auditors; **II) Partial Model Access:** Limited, structured model access (Bucknall & Trager, 2023) is granted within the secure environment, giving auditors visibility into key properties without complete exposure; **III) Secure Evaluation:** Standardized Evaluation Protocols (SEPs, see 3.2) for safety, bias, robustness, and alignment are executed within the protected environment; **IV) Cryptographically Signed Results**, to ensure their integrity (Ronis, 2024; Li et al., 2025); **V) Public Summary**: A concise, non-proprietary summary of findings is generated for public transparency, protecting sensitive details.

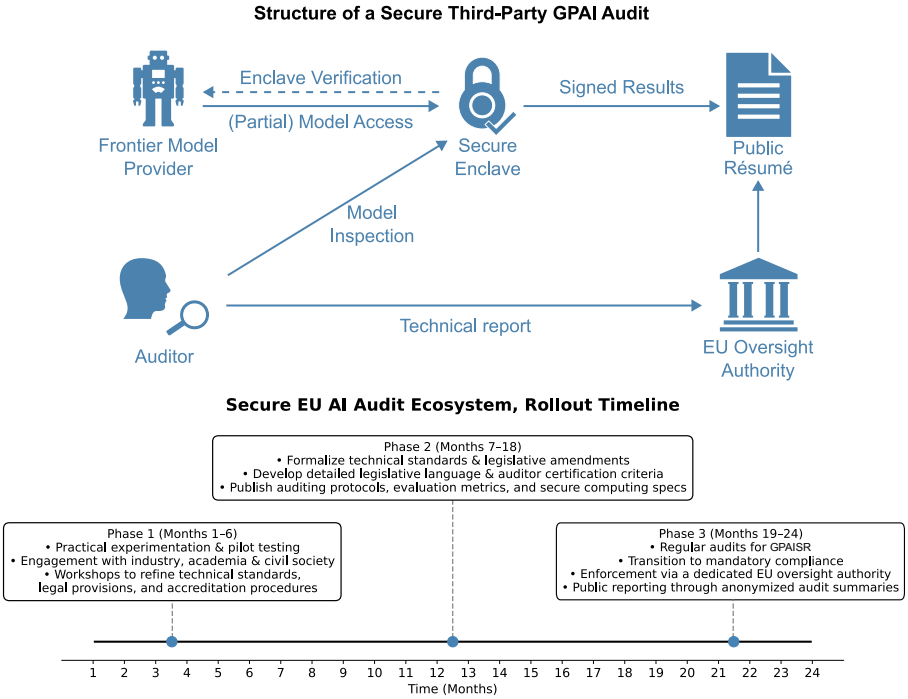
This architecture can be implemented through various technical approaches, including hardware-based TEE (Chrapek et al., 2024) such as Intel SGX or AMD SEV, software-based secure multi-party computation (Gamiz et al., 2025), or hybrid approaches combining black-box and white-box testing depending on security requirements (Islam, 2024; Tlaie & Farrell, 2025). Early community prototypes are already evaluating frontier models inside SGX clusters (OpenMined, 2025)

Practical limitations of current confidential computing. Process-level TEEs still ship with small dedicated secure memory (e.g. Intel SGX's 512 MiB EPC), but recent work shows that *full* Llama-2 7B and 13B can run inside SGX by paging most weights to encrypted DRAM, and inside VM-level Intel TDX whose protected memory scales to many GB (Chrapek et al., 2024). With careful quantisation and NUMA tuning these authors report $\sim$4–11% throughput and latency overheads relative to bare metal (far below the 2–3$\times$ figures often cited for naive paging). Nevertheless, end-to-end inference for *hundreds-of-billion* or trillion-parameter models still exceeds today's EPC/page-fault budgets. At the same time, there are encouraging results from early *GPU-level* TEEs, which are now commercially available on NVIDIA Hopper (H100/H200) GPUs. Benchmark results report $\sim$ 9% throughput overhead for up to 70 B models (Zhu et al., 2024). Finally, as side-channel vectors (cache, branch, power) also remain open research problems, so we continue to recommend a *split-compute* design: run the high-risk probes inside the enclave and keep bulk inference on external GPUs, reserving fully homomorphic encryption (FHE) for narrow, high-assurance checks.

For open-source models, we propose a modified framework that leverages transparency while still providing standardized evaluations. GPAISR open-source models would undergo pre-release "white-box" audits, while lower-risk models could opt for standard black-box evaluation for certification purposes.

3.1.1. Tailoring audits for open-source GPAISR

Open-source frontier models pose distinct opportunities and risks. To fully reap the former, we recommend addressing the latter through (i) **white-box pre-release audits** conducted before public weight release, focusing on emergent capability checks; (ii) mandatory publication of **reproducible build scripts** and **SPDX-style derivation logs** that capture full data/weight provenance; and (iii) ongoing **community red-teaming bounties** coordinated by the EU oversight body. These measures leverage transparency to



offset the absence of contractual NDAs while preserving competitiveness through timed releases.

3.2. Standardized Evaluation Protocols (SEPs)

Technical standardization is crucial for consistent evaluation across diverse AI architectures. We propose standardized test suites and detail different metrics; in brackets we signal whether it's desirable for them to be high (↑) or low (↓), as well as examples of currently existing benchmarks that could be used:

SEP governance. The initial SEP catalogue would be promulgated by the relevant EU authority and maintained by a multi-stakeholder Technical Standards Board comprising accredited auditors, GPAISR providers, academic experts, and civil-society observers. The Board would quarterly and issue an annual revision cycle, allowing rapid deprecation of obsolete benchmarks and adoption of new frontier-risk probes.

3.3. Institutional Support Structure

The technical infrastructure requires three supporting pillars:

Legal Framework and IP Protection: Robust contractual and regulatory safeguards protecting commercial interests while ensuring compliance with data protection requirements. Auditors would operate under strict non-disclosure agreements, legally required to handle sensitive information securely. This provides clear liability boundaries and creates legal certainty for all parties involved. Additionally, a parallel "public-interest access" tier would enable certified academic or civil-society researchers to perform limited audits under controlled conditions, fostering broader stakeholder engagement.

Secure Technological Standards: Beyond the core confidential computing technologies, we propose standardized protocols for secure information exchange, verified communication channels between model providers and auditors, and comprehensive audit logging mechanisms. These standards would define minimum requirements for both hardware and software components of the secure evaluation environment, ensuring consistent security properties across different implementations.

Accreditation and Oversight: A dedicated EU oversight function (either by expanding an existing body like DG CONNECT or establishing a new authority) would certify auditors, monitor compliance, and enforce standards. This authority would publish aggregated, anonymized audit results to enhance transparency while preserving commercial confidentiality. Over time, this authority could evolve into a standalone EU agency if scale or geopolitical considerations demanded further independence. We detail these two possible scenarios in Annex I. Cost estimation and funding instruments.

SEP governance. The initial SEP catalogue will be promulgated by the EU Audit Office and maintained by a multi-stakeholder Technical Standards Board comprising accredited auditors, GPAISR providers, academic experts, and civil-society observers. The Board meets quarterly and

SEP	Metric	Dataset	Pass/Fail Threshold	Evaluator role
Safety	Jailbreak success rate (↓)	1 000 HELM *Adversarial* prompts	$< 5\,\%$ critical-severity jailbreaks	Auditor executes prompts; provider supplies system policy
Bias	Average group-specific toxicity gap (↓)	Balanced DialBias	Gap < 0.05	Auditor
Robustness	Accuracy Δ under perturbations (↑)	AimBench robustness split	Drop $< 10\,\%$	Auditor
Alignment	Agreement with stated capability card (↑)	200 scenario-based vignettes	$\geq 90\,\%$ agreement	Joint: auditor validates provider-generated card

Table 1. Summary of Standardised Evaluation Protocols (SEPs).

issues an annual revision cycle, allowing rapid deprecation of obsolete benchmarks and adoption of new frontier-risk probes.

4. Implementation Roadmap and Key Challenges

4.1. Phased Technical Implementation

We propose a three-phase implementation strategy (see Fig. 3.1) to ensure technical feasibility and stakeholder acceptance. The audit ecosystem could be implemented through various institutional channels: **DG CONNECT Unit C1 (Cloud & Chips)**: Mobilize Chips Act facilities to support on-device secure auditing hardware development. **DG CONNECT Unit C3 (AI Factories)**: Integrate confidential audit modules within new AI manufacturing clusters. **EuroHPC Joint Undertaking**: Offer federated secure compute capacity for cross-border, confidential audit pilots. See Annex I. Cost estimation and funding instruments for a cost estimation of this implementation.

Phase 1: Technical Pilot serves as a rapid-learning sandbox where a small number of accredited auditors experiment with secure audits under controlled conditions. Technical teams would benchmark confidential computing approaches for performance and information leakage, assess secure access protocols, and test integration with existing model infrastructure. Pilot programs would include open-source AI projects on a voluntary basis to test transparency-compliance balances. Stakeholder workshops would refine technical standards, legal provisions, and accreditation procedures throughout this phase.

Phase 2: Formalization and Standardization builds upon the previous phase's insights to establish formalized protocols. This includes developing legal frameworks for mandatory audits, creating standardized APIs for model interaction, and specifying technical requirements for secure computing environments. Up to ten third-party auditing organizations would receive certification, with EU co-funding provided for specialized hardware and SME training. The specifications would include provisions for periodic expert-led reviews to adapt to emerging AI capabilities and security challenges.

Phase 3: Full Implementation transitions to structured compliance and operational maturity. Providers of AI models meeting systemic-risk criteria would demonstrate successful audit clearance before EU deployment, with the economic model shifting to private funding. The EU oversight authority would ensure compliance through scheduled and triggered audits, with re-audit protocols established for major model updates. The ecosystem would become a self-sustaining market where auditors compete on expertise and efficiency, while public accountability is maintained through anonymized audit summaries.

4.2. Technical Challenges and Solutions

Performance Overhead: Confidential computing introduces significant computational overhead that may impact evaluation efficiency. Research evaluating confidential computing with CPU-GPU configurations shows performance varies based on model types and batch sizes (Mohan et al., 2024), but overhead for large language models can be minimized to "below 5%, with larger models and longer sequences experiencing near-zero overhead" (Network, 2024). We propose hybrid approaches combining offline verification with selective runtime monitoring to minimize performance impacts (NVIDIA, 2023). Additionally, optimized secure computing protocols specifically designed for AI model evaluation could reduce the performance penalty associated with standard TEEs (Islam, 2024; Chrapek et al., 2024).

Information Leakage: Side-channel attacks and inference attacks may compromise the secure environment. TEEs are vulnerable to side-channel attacks that allow adversaries to learn secrets within enclaves, particularly through techniques that trigger exceptions or interrupts to trace the control or data flow. Advanced isolation techniques (Cui et al., 2023), formal verification of enclave security properties (Grimm et al., 2018), and differential privacy mechanisms will be necessary to mitigate these risks (Dwork & Roth, 2014). One promising approach involves "hardware-based security solutions such as secure enclaves and confidential

compute environments" which ensure sensitive computations remain isolated and encrypted even during processing (Global, 2025). The framework must specifically address model extraction risks (Carlini et al., 2024) through careful API design and access limitations (Bucknall & Trager, 2023).

Model Access Control: Determining the appropriate level of model access for meaningful evaluation without full disclosure requires careful technical design. A dynamic governance model should include a tiered certification system that rates AI systems based on their adherence to risk-mitigation guidelines (Media, 2024). We propose tiered access protocols with progressive disclosure based on findings, with different levels of access for different audit objectives (Tlaie & Farrell, 2025). This might include: **I)** Attention-mechanism visibility for bias audits; **II)** Gradient access for specific safety evaluations; **III)** Input-output behavior analysis for performance assessment; **IV)** Architecture details for specialized security audits (Casper et al., 2024).

Technical Standardization: Ensuring consistent evaluation across diverse model architectures requires architecture-agnostic testing focused on behavioral properties rather than implementation details. AI auditing can learn from established practices in financial accounting or safety engineering while acknowledging that academic researchers fill an important role by studying the feasibility and effectiveness of different AI auditing procedures. We propose establishing benchmarks for behavioral consistency, developing normalized evaluation metrics, and creating reference implementations for key audit procedures (Mökander, 2023).

Political Feasibility, Legal Harmonisation & Market Incentives: Divergent Member-State views on strategic autonomy may slow mandate adoption; phased pilot results translated into all official EU languages can build legitimacy. Also, Audit data flows must satisfy GDPRArt. 6(1)(f) legitimate-interest tests and dovetail with DSA Art. 37 audit obligations for VLOPs/VLOSEs to avoid duplicate reporting. Finally, a regulated audit-fee floor indexed to compute cost, coupled with InvestEU loan guarantees for SME auditors, balances entry incentives with long-run competition.

Rapid Technological Change: AI technologies advance rapidly, potentially outpacing rigid regulatory standards. Effective AI governance requires comprehensive auditing that is proportional to AI systems' capabilities and available affordances (Sharkey et al., 2024) working backward through the causal chain of effects (Demain, 2024). To address this, the EU would adopt flexible, modular standards subject to periodic expert-led reviews and updates (Sharkey et al., 2024). This ensures continued relevance as technologies evolve, keeping the auditing framework effective and resilient. Specialized "AI Horizon Scanning Units" embedded within the oversight authority could track emerging breakthroughs and recommend timely adjustments.

5. Conclusion

Our framework addresses the critical gap between regulatory intent and technical implementation in AI oversight. By enabling secure external auditing while protecting legitimate commercial interests, it provides a practical path forward for responsible AI governance.

We recommend immediate technical work to:

1. Benchmark performance of confidential computing approaches for AI auditing.
2. Develop standardized APIs for secure model interaction.
3. Create formal verification methods for audit environments.
4. Establish technical criteria for auditor accreditation.
5. Develop open-source–specific audit pipelines and SPDX-style provenance logs.

Rather than mandating new legal obligations immediately, the EC should authorize and support a coalition-of-the-willing to pioneer secure audit pilots under a formalized Code of Practice annex. This voluntary phase would benchmark technical feasibility, operational costs, and industry participation. Once critical thresholds are met, the same protocols could be transposed into binding Delegated Acts under the AI Act.

This technical foundation, coupled with appropriate legal and institutional structures, can create a sustainable ecosystem for AI oversight that balances innovation with accountability, positioning the EU as a leader in responsible AI governance.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Act, E. A. EU AI act compliance checker. 2024. URL https://artificialintelligenceact.eu/assessment/eu-ai-act-compliance-checker/.

Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku. *Anthropic*, 2024.

Anthropic. Claude 3.7 Sonnet and Claude Code. *Retrieved from https://www.anthropic.com/news/claude-3-7-sonnet*, 2025.

Boushey, G. The punctuated equilibrium theory of agenda-setting and policy change. In *Routledge handbook of public policy*, pp. 138–152. Routledge, 2012.

Bozzola, E., Spina, G., Agostiniani, R., Barni, S., Russo, R., Scarpato, E., Di Mauro, A., Di Stefano, A. V., Caruso, C., Corsello, G., et al. The use of social media in children and adolescents: Scoping review on the potential risks. *International journal of environmental research and public health*, 19(16):9960, 2022.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., et al. The malicious use of Artificial Intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018.

Bucknall, B. S. and Trager, R. F. Structured access for third-party research on frontier ai models: investigating researcher's model access requirements. In *Proceedings (Oxford Martin School, University of Oxford; Center for the Governance of AI)*, 2023.

Carlini, N., Paleka, D., Dvijotham, K. D., Steinke, T., Hayase, J., Cooper, A. F., Lee, K., Jagielski, M., Nasr, M., and Conmy, A. Stealing part of a production language model. *arXiv preprint arXiv:2403.06634*, 2024.

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., et al. Black-box access is insufficient for rigorous AI audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2254–2272, 2024.

Casper, S., Krueger, D., and Hadfield-Menell, D. Pitfalls of evidence-based AI policy. *arXiv preprint arXiv:2502.09618*, 2025.

Chrapek, M., Vahldiek-Oberwagner, A., Spoczynski, M., Constable, S., Vij, M., and Hoefler, T. Fortify your foundations: Practical privacy and security for foundation model deployments in the cloud. *arXiv preprint arXiv:2410.05930*, 2024.

Clark, J. and Hadfield, G. K. Regulatory markets for ai safety, 2025. URL https://arxiv.org/abs/2001.00078.

Cui, S., Li, H., Li, Y., Zhang, Z., Vilanova, L., and Pietzuch, P. Quanshield: Protecting against side-channels attacks using self-destructing enclaves. *arXiv preprint arXiv:2312.11796*, 2023.

Demain, P. Learning from history: Gpai serious incident reporting, 2024. URL https://www.pourdemain.ngo/en/post/learning-from-history-gpai-serious-incident-reporting.

Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Ferrer, X., Van Nuenen, T., Such, J. M., Coté, M., and Criado, N. Bias and discrimination in AI: a cross-disciplinary perspective. *IEEE Technology and Society Magazine*, 40 (2):72–80, 2021.

Gamiz, I., Regueiro, C., Lage, O., Jacob, E., and Astorga, J. Challenges and future research directions in secure multi-party computation for resource-constrained devices and large-scale computations. *International Journal of Information Security*, 24:1–29, 2025.

Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Global, V. Securing the future of ai: Defending against model exfiltration and side-channel attacks, 3 2025. URL https://www.ve3.global/securing-the-future-of-ai-defending-against-model-exfiltration-and-side-channel-attacks/.

Grimm, T., Lettnin, D., and Hübner, M. A survey on formal verification techniques for safety-critical systems-on-chip. *Electronics*, 7(6):81, 2018.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Islam, M. S. *Confidential Computing with Trusted Execution Environments*. PhD thesis, 2024.

Jones, B. D. et al. A general empirical law of public budgets: A comparative analysis. *American Journal of Political Science*, 53(4):855–873, 2009.

Li, K., Li, C., Yuan, X., Li, S., Zou, S., Ahmed, S. S., Ni, W., Niyato, D., Jamalipour, A., Dressler, F., et al. Zero-trust foundation models: A new paradigm for secure and collaborative artificial intelligence for internet of things. *arXiv preprint arXiv:2505.23792*, 2025.

Lundgren, M., Squatrito, T., and Tallberg, J. Stability and change in international policy-making: A punctuated equilibrium approach. *The Review of International Organizations*, 13(4):547–572, 2018.

Media, L. A dynamic governance model for ai, 2024. URL https://www.lawfaremedia.org/article/a-dynamic-governance-model-for-ai.

Mohan, A., Ye, M., Franke, H., Srivatsa, M., Liu, Z., and Gonzalez, N. M. Securing ai inference in the cloud: Is cpu-gpu confidential computing ready? In *2024 IEEE 17th International Conference on Cloud Computing (CLOUD)*, pp. 164–175. IEEE, 2024.

Mökander, J. Auditing of ai: Legal, ethical and technical approaches. *Digital Society*, 2(3):49, 2023.

Network, P. Confidential computing on nvidia h100 gpu: A performance benchmark study, 2024. URL https://phala.network/posts/confidential-computing-on-nvidia-h100-gpu-a-performance-benchmark-study.

Niemi, A., Sovio, S., and Ekberg, J.-E. Towards interoperable enclave attestation: Learnings from decades of academic work. In *2022 31st Conference of Open Innovations Association (FRUCT)*, pp. 189–200. IEEE, 2022.

Novelli, C., Taddeo, M., and Floridi, L. Accountability in Artificial Intelligence: what it is and how it works. *AI & Society*, 39(4):1871–1882, 2024.

NVIDIA. Protecting sensitive data and ai models with confidential computing, 12 2023. URL https://developer.nvidia.com/blog/protecting-sensitive-data-and-ai-models-with-confidential-computing/.

OpenMined. Secure enclaves for AI evaluation. https://openmined.org/blog/secure-enclaves-for-ai-evaluation/, 2025. Accessed 13 June 2025.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Ronis, J. Don't trust when you can verify: A primer on zero-knowledge proofs. 2 2024. URL https://www.wilsoncenter.org/article/dont-trust-when-you-can-verify-primer-zero-knowledge-proofs.

Sharkey, L., Ghuidhir, C. N., Braun, D., Scheurer, J., Balesni, M., Bushnaq, L., Stix, C., and Hobbhahn, M. A causal framework for ai regulation and auditing. *Publisher: Preprints*, 2024.

Shi, Z., Bergers, J., Korsmit, K., and Zhao, Z. Auditem: toward an automated and efficient data integrity verification model using blockchain. *arXiv preprint arXiv:2207.00370*, 2022.

Stix, C. Foundations for the future: institution building for the purpose of Artificial Intelligence governance. *AI and Ethics*, 2(3):463–476, 2022.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Tlaie, A. and Farrell, J. Securing external deeper-than-black-box GPAI evaluations. *arXiv preprint arXiv:2503.07496*, 2025.

Tramèr, F. and Boneh, D. Slalom: Fast, verifiable and private execution of neural networks in trusted hardware. *arXiv preprint arXiv:1806.03287*, 2018.

Tramèr, F., Kamath, G., and Carlini, N. Position: Considerations for differentially private learning with large-scale public pretraining. In *Forty-first International Conference on Machine Learning*, 2022.

Von Eschenbach, W. J. Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology*, 34(4):1607–1622, 2021.

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

Zhu, J., Yin, H., Deng, P., Almeida, A., and Zhou, S. Confidential computing on nvidia hopper gpus: A performance benchmark study. *arXiv preprint arXiv:2409.03992*, 2024.

Annex I. Cost estimation and funding instruments

The following cost scenarios align with the phased approach in 4.1. Specifically, the temporary public support in Phases 1–2 (for secure computing infrastructure, training grants, or pilot audits) is gradually scaled back so that, by Phase 3, the private sector assumes the bulk of the costs. This applies to both cloud-based and on-premises options. In either scenario, the cost of mandatory audits is ultimately borne by the audited entities themselves, with the EU's direct expenditure focusing on oversight, accreditation, horizon scanning, and incentives during the ecosystem's formative stages.

Overview of the EU AI Audit Ecosystem

The Secure EU AI Audit Ecosystem is a proposed framework enabling accredited third-party auditors to conduct deep technical evaluations of advanced AI systems under strict confidentiality. It emphasizes trusted computing environments (e.g. confidential computing with Trusted Execution Environments) so auditors can examine AI models without exposing proprietary data. Implementation is planned in three phases, from a small pilot to full-scale mandatory audits, with a dedicated oversight mechanism (either a new EU authority, such as was the case for the EU AI Office, or an expanded mandate for an existing body like DG CONNECT). Importantly, under our proposed funding model, the auditors themselves are not on the public payroll or embedded within government. Instead, the EU's role is to certify auditors and provide secure infrastructure and incentives in early phases, gradually transitioning all audit operations to a competitive private-sector ecosystem by Phase 3.

Assumptions: Each phase corresponds to a larger number of accredited auditing organizations and staff (scaling from a handful of audit teams in Phase 1 to dozens by Phase 3). Cost estimates below cover the major public expenditure components, secure infrastructure, legal and regulatory setup, accreditation systems, transparency portals, international coordination, and continuous technical updates, while **excluding ongoing auditor personnel salaries**, which in this model are financed by the audit organizations themselves by the later phases. These figures are rough, order-of-magnitude estimates for the initial rollout in each phase (all costs in Euros).

Phase-wise Cost Estimates by Infrastructure Scenario: The tables below present the estimated costs for each phase under two infrastructure scenarios (cloud-based vs. on-premises). Table 1 lists common expense categories that are essentially identical for both scenarios. Notably, the Personnel costs, by Phase 3 consisting on the $\sim$2,500 auditor Full-Time Equivalents (FTEs), are assumed to be paid by private audit firms or via fees from AI providers rather than by the EU budget. In place of those, a temporary support program in Phase 2 is included to help jump-start an SME-based audit industry (through one-time grants, training support, or compute credits). Table 2 shows the expense items that differ between the cloud-based and on-premise scenarios (primarily secure computing infrastructure costs). After showing both tables, we then provide narrative for each scenario's phase progression.

Scenario 1: cloud-based confidential computing infrastructure In this scenario, auditors utilize secure cloud services (confidential Virtual Machines, encrypted storage, etc.) to perform audits. This offers low upfront capital costs and flexibility, the EU can rely on pay-as-you-go cloud computing resources rather than invest heavily in dedicated hardware. It leverages commercial cloud providers (ideally European cloud platforms with built-in TEEs, although these most likely would be US-based providers) to minimize initial setup time. However, ongoing operational costs will scale with usage, and reliance on external providers means the EU must ensure data sovereignty and security compliance through careful configuration and contracts. Below we detail each phase under the cloud-based approach, highlighting the funding structure:

1. **Pilot (Phase 1):** This involves at most a handful of auditors in the pilot stage ($<$5 organizations with $\sim$10 FTEs each). Under a cloud approach, costs are relatively low, on the order of a couple of million euros, since no significant hardware is purchased. The majority of pilot expenses in a traditional model would be personnel, but in this approach those are not borne by the EU. Instead, the EU's Phase 1 spending focuses on setting up the secure cloud environment ($\sim$€ 1 M) and the necessary oversight and legal frameworks (e.g. NDA templates, basic guidelines, accreditation of a few initial auditors). Infrastructure costs are minimal by using existing cloud confidential computing instances for a few trial audits. Legal work in this phase covers drafting confidentiality agreements and ensuring pilot audits don't violate IP or data protection rules. Early accreditation efforts are basic (hand-picking a small number of qualified auditors to participate), and a simple reporting mechanism (e.g. a pilot audit summary) is set up. International outreach is limited to informing key partners about the pilot. Overall, Phase 1 costs in the cloud scenario are around $\sim$€ 2–3 M for the EU, reflecting the benefit of very low capital expenditure and no public outlay on auditor salaries.

2. **Formalization & Scale-Up (Phase 2):** Drawing on Phase 1 insights, Phase 2 lasts around one year and expands the ecosystem to $<$10 auditor organizations (up

Table 2. **Table 1: Common expenses for both scenarios (in €M)**

Cost Category	Phase 1	Phase 2	Phase 3
Legal & Regulatory Development	∼0.5	∼1–2	∼0.5/yr
Accreditation & Compliance Systems	∼0.2	∼1	∼0.5/yr
Transparency Interface	∼0.2	∼0.5	∼0.5/yr
International Coordination	∼0.3	∼0.5	∼1/yr
Horizon Scanning & Updates	∼0.1	∼0.3	∼1/yr
EU Co-Investment Windows (Phase 1–2)	-	∼10–15	-
Total	∼1–2	∼15–20	∼3–5/yr

Table 3. **Table 2: Scenario-specific expenses (in €M)**

Cost Category	Phase	Cloud-Based	On-Premises
Secure Computing / Infrastructure	1	∼1	∼2
	2	∼5	∼10
	3	∼0/yr	∼0/yr
Total (Common + Specific)	1	∼2–3	∼3–4
	2	∼20–30	∼25–35
	3	∼3–5/yr	∼3–5/yr

to ∼20 FTEs each, ∼200 total auditors engaged). This is the phase of standardizing and formalizing the audit framework. Personnel costs in the ecosystem would now climb into the tens of millions, as more auditors are active and the central oversight team grows, but critically, these salaries are largely paid by the emerging audit firms themselves rather than by the EU. The EU's budget in Phase 2 is directed toward supporting the nascent audit industry and building needed infrastructure. This includes developing a formal accreditation system (defining certification criteria, standing up an online platform for auditor accreditation and compliance tracking) and drafting detailed legislation and technical standards. Funds are allocated to expert workshops, documentation, and refining audit protocols and metrics for official use by the end of Phase 2. A more sophisticated public transparency portal is developed (to be launched in Phase 3) to disclose appropriate non-sensitive results and metrics. International cooperation efforts intensify (e.g. negotiating Memoranda of Understanding for cross-recognition of audit results). A key public expenditure in this phase is the "auditor ecosystem support" program: the EU may offer time-limited grants, subsidies for training, and free or discounted cloud compute credits to the new audit firms. These incentives ($\mathcal{O}(30M€)$) serve to lower the barrier to entry for Small and Medium-sized Enterprises (SMEs) specializing in AI auditing. For example, each accredited SME might receive a startup grant to hire expert staff or access specialized tools, and the EU could cover or discount their cloud computing bills for audit tests during this build-up period. Cloud infrastructure usage naturally increases in Phase 2 to support more frequent and complex audits, still with no upfront hardware investment, but higher cloud service fees commensurate with the greater volume of testing. The EU would likely cover a portion of these fees (as part of the compute credits incentive) to ensure firms can perform thorough evaluations of AI models without prohibitive cost. By the end of Phase 2, the EU's total spending in the cloud-based scenario is on the order of a few tens of millions of euros (approximately ∼€20–30 M), which is substantially lower than it would have been under a fully public model. This budget supports the creation of a competitive audit ecosystem while leaving the ongoing operating costs (especially auditor pay) increasingly to the private sector.

3. **Full Implementation (Phase 3):** By around month 18–24 of the rollout, the audit framework becomes mandatory for designated General-Purpose AI with systemic risk. Up to ∼50 accredited auditor organizations (∼50 FTEs each) may operate, creating a capacity of ∼2,500 auditors across the EU. Under full implementation, the auditing workforce becomes the dominant cost factor, the ecosystem requires a large, skilled workforce (AI safety experts, cybersecurity analysts, etc.) to conduct continuous audits, plus a suitably staffed oversight authority to enforce compliance. In a traditional public-funded model, this could scale to ∼100s M annually in human resources if all auditor po-

sitions were publicly funded. However, in our proposed model these personnel costs are borne by the private market (audit firms recouping costs via fees charged to AI providers for mandatory audits). The EU's direct role in Phase 3 is primarily regulatory and facilitative: maintaining the legal and standards framework, keeping the accreditation and transparency systems running, and coordinating internationally, tasks which together we estimate to cost only on the order of a few million euros per year. The cloud infrastructure spend also grows substantially in Phase 3 as hundreds of audits are performed (including very compute-intensive deep evaluations of frontier models), potentially amounting to tens of millions of euros per year in cloud computing and storage fees. Under this model, those cloud usage costs too would be paid by the audit firms or by audited companies, not by the public sector (beyond any minimal support in edge cases). By Phase 3, the EU would not be subsidizing routine audits: cloud computing resources for audits are procured by the auditors or passed through as part of their service fees. Other cost components stabilize at their operational levels, the legal/regulatory framework is largely in place (only minor updates and enforcement actions need funding), the accreditation and reporting systems are fully functional (with ongoing maintenance updates), and continuous horizon-scanning ensures technical standards keep up with AI advances. The EU oversight authority engages in active international agreements to mutually recognize audit results and avoid duplication of efforts across borders. In total, the EU's direct expenditure in Phase 3 under the cloud approach is only on the order of ∼€3–5 M per year, mainly for oversight and coordination. The overall ecosystem cost remains much higher (roughly ∼€120–150 M/yr when including all auditors and compute), but that is at this point financed by industry through the established audit fee structure. In other words, by Phase 3 the audit ecosystem is fully self-sustaining privately, with the EU providing regulation, certification, and minimal ongoing support rather than funding operations.

Scenario 2: on-premises secure facilities This scenario assumes building dedicated on-premise secure computing facilities (e.g. EU-managed data centers or auditor-owned secure hardware) for conducting the audits. This would align well with recent initiatives[1][2], so it could be a part of these. Here, upfront investments are higher, the program must purchase and set up hardware, secure enclaves, and physical infrastructure, but once in place, operational costs for compute may be lower per use (since auditors own or have dedicated access to the equipment). This approach gives auditors and regulators full control over security (sensitive models and data never leave the premises, addressing confidentiality concerns more directly). However, it requires more lead time and capital in the early phases to deploy trusted hardware and secure data facilities. Under the proposed funding model, the EU would still invest in these secure facilities initially, but the auditors' organizations would gradually take over ownership and operational costs. In principle, the EU may also charge competitive prices for those using these facilities, mitigating in part the initial investment. Most non-infrastructure activities and costs in the on-premises scenario remain similar to the cloud scenario (e.g. the number of people involved, the legal and administrative efforts, etc.). The difference lies primarily in how the computing environment is provided and who bears those costs over time:

1. **Pilot (Phase 1):** In Phase 1, instead of renting cloud instances, the program procures and configures a small secure computing environment for the pilot, for example, a limited-capacity data center module or a mobile secure lab with hardware TEE support. This upfront capital expense (approximately ∼€2 M) makes the Phase 1 cost slightly higher in the on-prem scenario (∼€3–4 M total) compared to the cloud approach. As in Scenario 1, the EU does not pay for auditors' salaries in the pilot; the few pilot audit teams are either self-funded (e.g. via existing research grants or company contributions) or compensated through some other mechanism. The EU's budget in Phase 1 instead covers the infrastructure purchase and setup, along with the same legal and coordination tasks described in Scenario 1 (drafting NDAs, basic oversight staff, etc.). The secure pilot facility ensures that during initial audits, data and models remain on hardware fully under EU or auditor control. At this stage, the capacity is limited (sufficient for a handful of audits), and the focus is on proving the concept of third-party audits in a contained, secure environment. Phase 1 on-premise costs are still within single-digit millions and primarily driven by the one-time hardware setup, since personnel costs are minimal for the EU (auditors are not on the payroll).

2. **Scale-Up (Phase 2):** By Phase 2, the audit ecosystem is formalized and grows to $<$10 auditor organizations (∼200 total auditors across them, similar to the cloud scenario). Significant capital investment is made to equip these auditor organizations with trusted hardware or to establish shared secure facilities they can access. For example, the program might fund a few regional secure computing centers with high-assurance

[1] https://digital-strategy.ec.europa.eu/en/policies/ai-factories

[2] https://ec.europa.eu/commission/presscorner/detail/en/ip_25_467

hardware for audits, or provide grants for each auditor firm to buy compliant secure servers. In practice, this means the EU dedicates on the order of ∼€10 M in Phase 2 for secure computing infrastructure, assuming co-funding of hardware at several sites, rather than the EU covering 100% of all equipment. This could involve matching funds or equipment grants where the audit firms also invest some of their own capital. In addition, the EU might facilitate low-interest loans or guarantees (through instruments like InvestEU) to help private audit providers finance their portion of the hardware procurement. Aside from infrastructure, the other activities in Phase 2 mirror those in the cloud scenario: building out the accreditation and compliance systems, refining standards and legislation, developing the transparency portal, and scaling up training and coordination. The EU would also run the temporary support program here, though focused less on cloud credits and more on direct grants and training, to ensure the new audit companies can hire and train staff and make use of the new secure facilities. Some funding may go toward specialized technical training for auditors on the on-prem systems, and ensuring interoperability and secure interconnects between different audit nodes (potentially leveraging programs like CEF Digital for cross-border infrastructure integration). By the end of Phase 2, the on-premises approach would have higher cumulative public expenditure than the cloud approach, due to these capital outlays, roughly ∼€25–35 M in EU spending in Phase 2 (including infrastructure and other support). The hardware deployed in Phase 2 provides a foundation that can be expanded or reused in the next phase. Official audit protocols and standards would be published by this point, and the audit firms should be operationally ready for mandatory audits.

3. **Full Implementation (Phase 3):** By Phase 3, the on-premises infrastructure needs to scale up dramatically to handle EU-wide audit demand when the audit requirement becomes mandatory for all relevant AI systems. Supporting ∼50 audit teams across Europe with dedicated secure computing means potentially deploying tens of millions of euros in additional hardware, secure data center space, and support systems to reach full capacity. Under the proposed model, the expectation is that this Phase 3 expansion of infrastructure is financed by the private sector and other non-EU sources rather than new EU grants. Audit firms that established themselves in Phase 2 would acquire additional hardware (possibly using their own capital or financing from commercial loans or national programs) to meet the demand. The EU may facilitate this by brokering agreements or standards for interoperability, and Member States might contribute via their own initiatives (for example, using Recovery and Resilience Facility funds to build local audit data centers, as envisaged in the original funding instruments). The hardware deployed in Phase 2 is expected to be reused and scaled, many audit organizations will build on the equipment they already co-funded, adding capacity as needed. By this stage, the EU ceases direct subsidies for audits: no routine operational funding is provided for either personnel or infrastructure. The ongoing public costs are limited to maintaining the oversight authority and its activities (regulatory updates, international coordination, and ensuring the audit process runs smoothly across the network). The auditors' own operational costs (staff, equipment maintenance, electricity for data centers, etc.) are covered by the fees they charge to AI providers for audit services. After the Phase 2 build-out, the annual infrastructure-related costs for audits become mostly maintenance, e.g. power, cooling, security and IT maintenance, which are typically lower than equivalent cloud rental costs for the same capacity. Those maintenance costs, too, would be borne by the audit firms or passed on to clients, not by the EU. In summary, by Phase 3 the on-premise approach yields a situation where the EU's direct spending is again only on the order of a few million euros per year (for oversight personnel and continued governance tasks), comparable to the cloud scenario.

Cost comparison and trade-offs

The choice between cloud-based and on-premise infrastructure significantly shapes the cost structure, scalability, and risk distribution in the audit ecosystem. Cloud solutions, especially leveraging confidential computing on established platforms, are cheaper and faster to deploy initially. They avoid large capital expenditures and offer pay-as-you-go flexibility, which is particularly useful in early phases when audit demand is uncertain and scaling needs to be responsive. In the context of a privately run audit ecosystem, the cloud approach lowers barriers to entry for new audit firms: an SME can start offering audit services without needing upfront capital for a data center, instead using cloud resources (the costs of which can be gradually covered by revenues or temporary EU credits). This flexibility facilitates rapid growth of the ecosystem. Cloud-based auditing does, however, entail ongoing operational expenses that track with usage. In the new model, those expenses become the responsibility of private actors, meaning audit firms must recuperate cloud costs through the fees they charge. The EU might need to oversee cloud provider arrangements to ensure data sovereignty and security (e.g. that European data protection standards are met on whatever cloud is used) and possibly to bulk-negotiate pricing or provide initial credits, but it does not need to budget for continuous cloud payments beyond

Cloud-Based Infrastructure	On-Premise Secure Facilities
Cost Structure: Low upfront costs, pay-as-you-go; higher recurring fees over time.	**Cost Structure:** High upfront capital costs; potentially lower long-term operational costs.
Scalability: Highly scalable; quick to deploy; flexible for changing demand.	**Scalability:** Slower to scale; requires hardware procurement and setup time.
Security/Control: Relies on external providers (possibly US-based); strong encryption possible.	**Security/Control:** Full control over data and hardware; more suitable for highly sensitive audits.
Operational Burden: Minimal IT overhead for auditors; provider handles infrastructure.	**Operational Burden:** Higher complexity; requires dedicated IT, maintenance, and physical security.
Best Use Case: Ideal for pilot and early phases; lowers barriers to entry.	**Best Use Case:** Suitable for mature, high-volume, high-trust deployments.

Table 4. Comparison of Cloud-Based Infrastructure (Scenario 1) vs. On-Premise (Scenario 2)

Phase 2.

On-premise infrastructure, by contrast, requires significant upfront investment in secure hardware, facilities, and IT personnel. In a fully public scheme this upfront cost is a burden on the state, but in the revised approach it becomes a shared burden: the EU helps kick-start the infrastructure in early phases, and private audit firms (along with possibly Member States or financial institutions) take on the capital expansion later. This can lead to more sovereign control over the audit process, data never leaves local premises, and auditors are not dependent on third-party cloud providers, which is advantageous for highly sensitive audits and for assurance of compliance. Over the long run, if the hardware is efficiently utilized, on-premise solutions might yield lower marginal costs per audit (since owning equipment can be cheaper than renting equivalent cloud capacity over many years). Those savings would accrue to the audit firms or their clients, as they would no longer pay recurring cloud fees once the equipment is paid off. From the EU's perspective, the on-prem model front-loads the need for support (Phase 1–2 grants for hardware) but then allows public spending to taper off completely by Phase 3, as the private sector assumes both operational and capital expenditures.

In summary, under this phased, private-sector-driven funding model, both scenarios converge by Phase 3 in that the EU's direct financial involvement is minimal and the audit ecosystem is self-sustaining. The cloud-based approach offers a faster, lower-cost ramp-up for the EU and the auditors in Phase 1–2, while the on-premises approach demands more initial investment (from public and private sources) to potentially reap benefits in security and long-term cost efficiency. Policymakers will need to balance these trade-offs.

We suggest that the EU could pursue a hybrid strategy: using cloud infrastructure in the pilot and scale-up phases to grow the ecosystem quickly, while encouraging or co-financing on-premise capacity for the long run. Crucially, the funding instruments to support either choice are available, e.g. Digital Europe Programme grants for infrastructure and training, Horizon Europe for audit tool R&D, InvestEU for loans to audit firms, and CEF Digital for cross-border connectivity. These can be applied in Phase 1–2 to ensure the ecosystem reaches critical mass. By Phase 3, however, the goal is that audits of GPAISR are a normal market service: audit firms compete and innovate, AI providers pay for compliance (just as companies pay for financial audits), and the EU maintains oversight and facilitates the ecosystem's continued growth without direct operational subsidies. This phased approach secures the same overall capacity for trustworthy AI auditing, roughly 50 organizations and 2,500 expert auditors in steady state, but with a sustainable financing model that leverages private sector efficiency and entrepreneurship.

Oversight body: new EU authority vs. expanding DG CONNECT

Institutional design will also affect both effectiveness and cost. Embedding the oversight function within an existing body like DG CONNECT offers a fast, lower-cost option. It allows the EU to begin coordinating pilot audits, developing standards, and certifying auditors without creating new administrative structures. This approach is well-suited for the early phases, where agility and low friction are essential.

However, as the audit framework matures and the number of participating auditors increases, a standalone EU authority may be necessary to ensure independence, sustained focus, and credibility. A dedicated agency would be better equipped to enforce compliance, manage sensitive information, and serve as a central point for international coordination. Although more expensive and slower to establish, it could mirror successful models like ENISA or the European Data Protection Board.

A pragmatic path would begin with an audit unit inside DG CONNECT during Phases I–II, followed by a transition to an autonomous oversight body in Phase III once the system is proven and institutional momentum has been established.

Annex II. Alternative potential policies for the EU AI Auditing ecosystem

This section sets out a three-tiered policy approach (**Step**, **Jump**, and **Leap**, SJL) to guide EU-wide measures for establishing an ecosystem of third-party auditors of general-purpose AI (GPAI) models with systemic risk. Although no formalized SJL structure exists in current technology policy literature, we propose it as a useful tool for aligning policy ambition with the evolving nature of AI risks and governance challenges.

Technology policy exhibits some key features that suggest that such a tiered approach may be useful:

1. Technology policy often advances in small increments until major events (e.g., notable failures or dramatic advancements) trigger rapid regulatory adjustments (Jones et al., 2009; Boushey, 2012; Lundgren et al., 2018). AI's evolution demonstrates that steady progress can be punctuated by breakthroughs that reshape our understanding of risks and the necessary oversight. Thus, a tiered system permits rapid escalation from modest, incremental measures (**Step**) to transformative actions (**Leap**) when warranted.
2. Especially for cutting-edge AI systems, the full scope of risks may become apparent only after widespread deployment (Brundage et al., 2018). As a result, governance is often reactive: by the time evidence of potential harm surfaces, the technology might be well entrenched (Bozzola et al., 2022). Consequently, some authors have recently argued against evidence-based policymaking (Casper et al., 2025) in the context of AI Risk.
3. Policymakers must avoid both extremes: being too lenient, thereby permitting risky deployments, or too restrictive, which could come with an opportunity cost. A tiered policy approach enables regulatory interventions to be tailored to the evolving evidence of risk.
4. Different policy proposals require varying degrees of resource investment, technical sophistication, and political capital. Simpler interventions generally face lower resistance but may be insufficient for managing systemic risks, while more robust measures need substantial investments and stronger political consensus.

Together, these features support the categorization into **Step**, **Jump**, and **Leap** tiers. Each category reflects an increasing degree of intervention, political commitment, and resource allocation, ensuring that the EU's response to frontier AI challenges can be scaled appropriately to the severity and urgency of emerging risks.

Applying the SJL to the Secure EU GPAI Auditing ecosystem

Below, specific policy measures are detailed for each tier. Although the approaches are presented as discrete categories, they can be implemented incrementally or in combination, thereby ensuring a flexible yet robust framework for regulating AI auditing in the EU. In 2, we score all of the policy initiatives on different axes (*Cost*, *Norm-Setting Potential*, *Political Feasibility*, *Speed*, *Technical Maturity*, and *Stakeholder Buy-In*) in all cases, we have set the score up in such a way that "higher is better". In 1, we then summarize these scores when grouping initiatives into **Step**, **Jump**, and **Leap** tiers.

STEP: INCREMENTAL, LOW-COST, LOW-POLITICAL WILL

Step measures are modest in scope and are typically introduced via existing policy channels and legal frameworks. Their aim is to enhance transparency and foster best practices without imposing significant new burdens.

- Publish technical **audit guidance via an AI Act Code of Practice annex:** Policymakers would issue baseline guidelines to clarify recommended audit procedures, metrics, and documentation standards, helping developers align with emerging best practices voluntarily.
- **Develop a "Transparency Toolkit"** for developers: This resource would include template model cards, logs of fine-tuning datasets, and recommended safety-evaluation protocols. The goal is to foster consistent, transparent documentation practices—particularly useful for smaller organizations.
- **Encourage voluntary "Audit-Ready" labels** for AI providers: Organizations could self-attest to meeting certain pre-audit criteria (e.g., detailed version control, robust logging). While not mandatory, these labels bolster reputational standing and signal readiness for more formal audits if needed.

	New EU Oversight Authority	Expanded DG CONNECT Unit
Setup Speed	Slower; requires new legislation and administrative setup.	Faster; leverages existing structures.
Costs	Higher fixed costs (staff, offices, admin overhead).	Lower marginal cost; builds on current resources.
Expertise and Focus	Mission-specific, tailor-made institution.	Risk of diluted focus amid other DG priorities.
Autonomy & Legitimacy	Greater independence, possibly more public trust.	More aligned with EU Commission policy oversight.

Table 5. Comparison of a New EU Oversight Authority vs. an Expanded DG CONNECT Unit

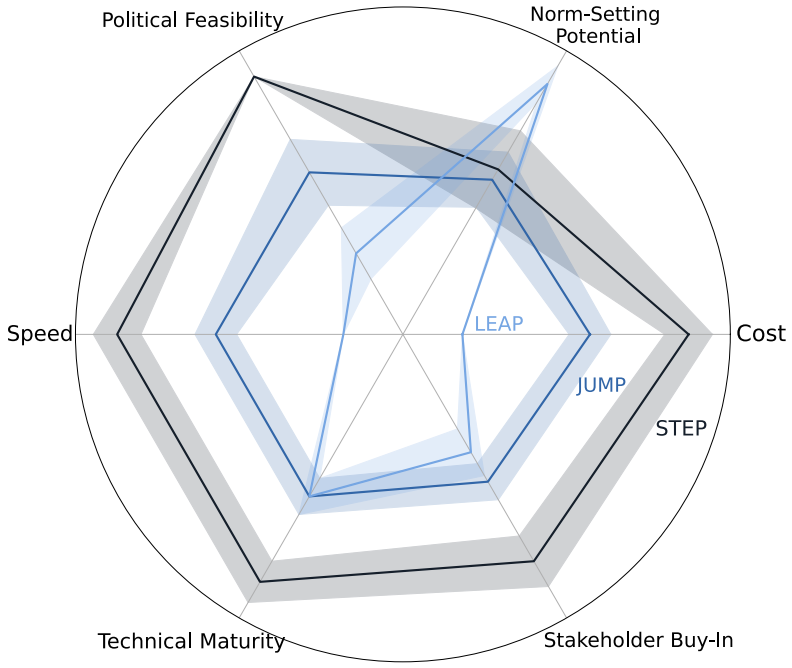


Figure 1. **Average policy interventions scores.** Across categories (Step, Jump and Leap), average the score (solid line) of different policy interventions (shaded region: $\pm 1\sigma$) across dimensions. Higher is better.

- **Clarify GDPR compatibility for audit data access** via new EDPB guidance: Official guidance would delineate how sensitive information can be processed or shared for audit purposes under EU data protection laws, enabling more effective—but still privacy-compliant—voluntary audits.

- **Develop shared audit test suites** with open participation: In partnership with industry and civil society, regulators would facilitate open benchmarking platforms, allowing multiple stakeholders to test AI systems using common audit tools and methodologies.

These **Step** interventions serve as a foundational layer—minimally intrusive measures that can be scaled up as needed.

JUMP: MODERATE COST AND POLITICAL WILL, TARGETED INTERVENTIONS

Jump policies involve more formal and resource-intensive measures, emerging when voluntary actions are insufficient for mitigating risk. They typically include the establishment or reinforcement of oversight institutions.

- **Launch a "Trusted Auditors Network"**, with EU-backed accreditation and pooled secure compute access: Independent AI auditors would be formally accredited under a common framework, with access to EU-supported infrastructures (e.g., secure data centers) to conduct in-depth evaluations of GPAISR models. This would constitute Phases I and II of our main policy proposal.

- **Expand DG CONNECT with a Dedicated Audit Oversight Taskforce** (precursor to a new agency): This specialized unit would coordinate audits, monitor emerging risks, and suggest regulatory updates, effectively centralizing audit expertise at the EU level. This is one of the scenarios we discuss in *Annex I. Cost estimation and funding instruments*.
- Introduce technical **annexes to the AI Act that codify audit format, scope, and risk thresholds:** Detailed requirements or guidelines under the AI Act would standardize how audits must be conducted, ensuring consistency across Member States.
- Define categories of systemic GPAI that require **mandatory confidential pre-deployment audits:** Advanced or globally impactful AI systems would be subject to independent, third-party risk assessments before market entry, mitigating potentially systemic threats. This would be the equivalent of Phase III of our main policy proposal.
- **Fund a public "Audit Compute Commons":** EU-hosted secure environments (e.g., TEE/cloud VMs) would be made available to accredited auditors and SMEs, lowering financial and technical barriers to rigorous AI evaluations. This is one of the main scenarios we detail in *Annex I. Cost estimation and funding instruments*.
- **Establish a legal sandbox for open-source model audits** (non-profit access with confidentiality protections): Non-profit entities and research teams could obtain special protections for conducting open-source AI audits, balancing transparency with intellectual property considerations.

These interventions would add layers of accountability and oversight without curtailing AI innovation, effectively bridging the gap between minimal and transformative regulatory measures.

LEAP: HIGH-COST, HIGH-POLITICAL WILL, TRANSFORMATIVE MEASURES

Leap policies represent a fundamental transformation of the regulatory framework and are reserved for situations where AI risks are severe or systemic.

- **Create a standalone EU AI Auditing Authority:** Similar in structure to ENISA or the EDPS, this body would possess legal powers to accredit auditors, sanction non-compliance, and enforce market restrictions on unsafe AI systems. This is one of the main scenarios we detail in *Annex I. Cost estimation and funding instruments*.
- **Codify a "GPAI Systemic Risk" designation scheme**, requiring mandatory pre-release audit sign-off: Inspired by REACH (chemicals regulation) [3], advanced AI models flagged as "systemic risk" would undergo formal approval steps prior to deployment.
- Integrate **confidential computing as as the default for auditing GPAISR models:** As discussed in the main text, leveraging secure enclaves or multi-party computation, sensitive data and intellectual property remain protected during thorough third-party audits.
- **Use trade leverage** (e.g., Digital Markets Act + AI Act) to require audit reciprocity for market access: The EU could condition access to its large internal market on adherence to equivalent audit standards in partner countries. This could be thought of as a "Brussels Effect for AI Audits".
- **Offer equivalence treaties to partners** (US, UK, Japan) who adopt compatible auditing norms, creating a Global GPAI Audit Accord: Beyond reciprocal market access, this fosters a unified international framework for AI oversight and trust.
- **Launch a "European Audit Compute Facility"**, similar to EuroHPC but tailored for secure model audits: This infrastructure would host large-scale computing resources for independent evaluations of frontier or systemic AI, overseen by EU institutions. This is an extension of one of the scenarios we detail in *Annex I. Cost estimation and funding instruments*, and it would be aligned with the recently proposed AI (Giga)Factories.
- **Mandate standardized post-audit disclosure reports**, akin to financial disclosures, overseen by the new audit authority: GPAISR developers would regularly submit uniform audit summaries, ensuring ongoing regulatory insight into performance and safety.

Although these transformative measures demand substantial resources and political capital, they provide the most robust safeguard against catastrophic or systemic AI risks by fundamentally overhauling the regulatory landscape.

In sum, structuring the EU's approach to AI auditing using **Step**, **Jump**, and **Leap** categories provides a coherent framework that calibrates regulatory action in line with the dynamic, evolving risk landscape. Starting with pragmatic, low-cost actions and reserving more comprehensive interventions for cases where systemic risks can be reasonably anticipated, EU policymakers can develop a governance framework for frontier AI that remains both adaptable and robust.

[3] https://environment.ec.europa.eu/topics/chemicals/reach-regulation_en

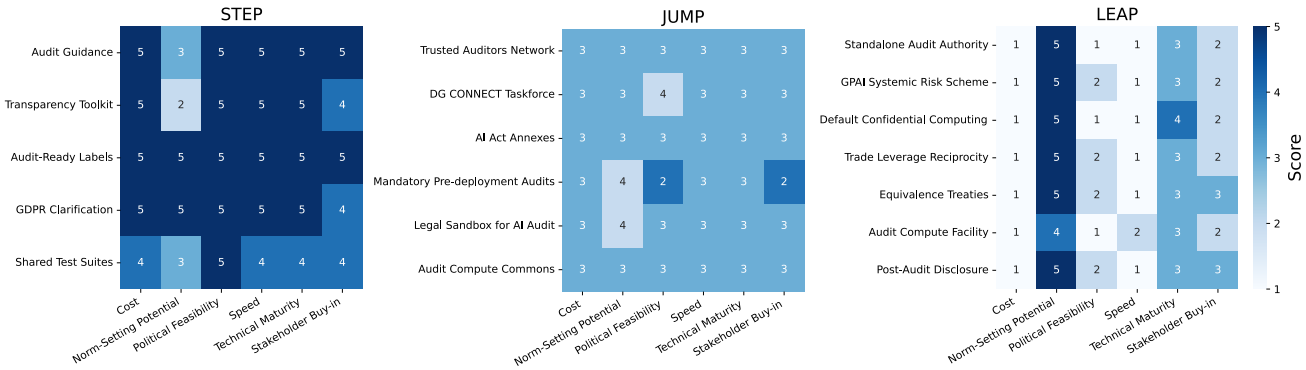


Figure 2. **Policy interventions scores.** Across categories (Step, Jump and Leap), we (subjectively) score each policy intervention across different dimensions. Higher is better.