Private Compute Permit Markets for AI Assurance: An Incentive-Compatible Architecture That Links Evaluations, Insurance, and Procurement

Joel N. Christoph
European University Institute
joel.christoph@eui.eu

Abstract

Private governance can complement public regulation by shaping incentives for responsible AI development. This tiny paper proposes a concrete market design for risk-calibrated compute permits that integrates three private oversight tools: independent evaluations, insurance with capital at risk, and procurement clauses. The design defines a permit unit indexed by capability-adjusted compute, a clearing mechanism with supply caps tied to risk budgets, and a settlement workflow where evaluation providers act as oracles whose scores parameterize insurer premiums and performance bonds. Developers must hold permits in proportion to capability-adjusted compute for training or deployment, while purchasers require proof of permits and insurance at award. A simple model shows that premiums increasing in measured risk induce safer effort when developers face permit constraints and assurance pricing. We outline a minimal attestation schema that clouds, insurers, and registries can implement without exposing proprietary artifacts, and we propose a 12 week pilot with one cloud, two evaluation providers, and an insurer. This design connects technical evaluations to market incentives, offering a practical private governance mechanism that is compatible with varied regulatory contexts.

1 Motivation and contribution

Private governance mechanisms such as evaluations, insurance, and procurement are gaining traction as complements to regulation. We contribute an implementable architecture that links these tools through a compute permit market. The proposal is designed for rapid piloting and for integration with existing assurance workflows.

Contributions:

- A definition of a capability-adjusted compute permit unit and a corresponding issuance and clearing mechanism.
- An oracle design that maps standardized evaluation scores to insurance pricing and permit requirements.
- A procurement clause template that requires proof of permits and insurance at contract award and at renewal.
- A minimal attestation and audit schema that preserves confidentiality while enabling verification.

2 Mechanism design

Actors: developers D, evaluation providers E, insurers I, cloud providers C, a permit registry R, and purchasers P.

Permit unit: one permit covers a quantity of compute Q scaled by a risk multiplier $m \in [1, M]$ derived from standardized evaluations. Effective compute $Q^* = mQ$. Developers must retire Q^* worth of permits for training or deployment events above a threshold.

Issuance and clearing: the registry R issues a capped supply of permits per period. Permits are auctioned to developers and brokers. Unused permits expire. Caps can be set by an independent governance board advised by E and I.

Evaluation oracle: providers E publish signed reports with a composite capability and safety score. Scores instantiate m via a public mapping. Reports reference test suites and conditions, enabling comparability across providers.

Insurance and bonds: insurers I price coverage and performance bonds as a function of m, incident history, and controls. Coverage includes first party loss, third party liability, and incident response. Bonds are forfeited on verifiable non-compliance.

Procurement: purchasers P include a clause that requires proof of permits and valid insurance before award and at each major model update. Clouds C provide permit metering hooks and attestation logs.

3 A simple incentive result

Let developer effort $e \in [0, 1]$ reduce evaluated risk so that $m = \bar{m} - \alpha e$ with $\alpha > 0$. The developer chooses e and scale Q to maximize

$$\max_{e,Q} \ \Pi(Q) - p(m)Q - \lambda mQ - c(e),$$

where $\Pi(Q)$ is revenue, $p(m)$ is the insurance premium per unit compute, λ is the shadow price of permits, and $c(e)$ is convex cost of effort. If $p'(m) > 0$ and permits bind with $\lambda > 0$, the first order condition in e implies $c'(e) = (\alpha)(p'(m) + \lambda)Q$, so higher evaluated risk raises the marginal return to safety effort. Hence evaluations coupled to insurance pricing and permit constraints create aligned incentives for safer development.

4 Implementation sketch

Attestation schema: each training or deployment event emits a signed JSON object with fields `model_id`, `commit`, `eval_report_hash`, `effective_compute`, `permit_ids`, `insurer_policy_id`, and `timestamp`. Registries validate signatures and retire permits.

Pilot plan: run a 12 week pilot with one cloud for metering, two evaluation providers for the oracle, and one insurer. Success criteria include end-to-end attestations, consistent pricing across providers, and purchaser acceptance of the proof bundle.

5 Limitations and open questions

Evaluation coverage can be incomplete, and permit caps can be mis-specified. Cross-jurisdictional interoperability and avoidance of regulatory arbitrage require governance safeguards. Further work should test robustness to gaming and measure real-world safety effort responses.

Broader impacts

If successful, the approach can direct investment and engineering effort toward risk reduction while preserving innovation. Poor implementation could create artificial scarcity or entry barriers, so safeguards and open standards are important.

References

- Hadfield, G., Clark, J. 2023. Regulatory Markets: The Future of AI Governance. arXiv.
- Tomei, G., et al. 2025. AI Governance through Markets. arXiv.
- Ball, R. 2025. A Framework for the Private Governance of Frontier AI. arXiv.
- Liang, P., et al. 2022. HELM: Holistic Evaluation of Language Models. arXiv.
- Mazeika, M., et al. 2024. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming. arXiv.
- Mitchell, M., et al. 2019. Model Cards for Model Reporting. FAccT.
- Gebru, T., et al. 2021. Datasheets for Datasets. CACM.
- Kirchenbauer, J., et al. 2023. A Watermark for Large Language Models. NeurIPS.
- Pruthi, G., et al. 2020. Estimating Training Data Influence by Tracing Gradient Descent. NeurIPS.
- Weil, G., et al. 2024. Insuring Emerging Risks from AI. University of Oxford.