The Inference Gap: Test-Time Compute Scaling and the Limits of Training-Centric AI Governance

Luiz Felipe Caparelli Piochi [1]

Abstract

Contemporary AI governance frameworks, including the EU Artificial Intelligence Act and frontier developers' responsible scaling policies, assume that AI capability is determined primarily at training time. Test-time compute (TTC) scaling, wherein the same model weights produce qualitatively different outputs depending on inference budget, exposes gaps in this assumption that existing frameworks do not address. We identify three governance gaps that result from this mismatch. The first is a *threshold decoupling gap*, through which developers can pair sub-threshold training runs with inference-time amplification to achieve frontier-level capabilities outside the scope of enhanced regulatory scrutiny. The second is an *evaluation staleness gap*, through which pre-deployment safety assessments conducted at fixed inference budgets fail to characterise capabilities that emerge under extended reasoning. The third is a *reasoning opacity gap*, through which hidden chain-of-thought processes undermine the observability requirements that auditable governance depends on, with the additional complication that visible chain-of-thought is not reliably faithful even where accessible. We further analyse how agentic deployment compounds all three gaps and propose a set of technically grounded reforms, including inference-budget-aware capability evaluations and reasoning trace logging for regulated applications.

1. Introduction

The rapid proliferation of AI governance frameworks across the European Union, the United States, the United Kingdom, and beyond reflects a growing recognition that the societal implications of AI development require structured oversight (Reuel et al., 2024; Trager et al., 2023). These frameworks differ substantially in scope and enforcement, but they share an assumption that training compute is a sufficient proxy for the deployment-time capability profile that governance seeks to regulate. The EU AI Act, for instance, imposes Article 72 post-market monitoring obligations on high-risk systems (European Parliament and Council, 2024), and developer-led responsible scaling policies trigger re-evaluation at capability milestones. Our claim is narrower and more specific, pointing that the principal regulatory triggers, conformity assessments, and capability evaluations are designed around the trained artifact and a single inference regime, and they do not treat the per-call inference configuration as a governable parameter. Test-time compute (TTC) scaling exposes this design choice as the load-bearing premise it is.

This assumption was sound when training compute was the dominant capability driver. Empirical scaling laws confirm a predictable relationship between training compute and performance, making it a tractable governance proxy. The EU AI Act uses cumulative training FLOPs as its highest risk criterion, while US Executive Order 14110 required reporting triggered by training compute thresholds (European Parliament and Council, 2024; White House, 2023). This approach connects an observable, auditable quantity to governance targets.

A growing body of research demonstrates, however, that this mapping is no longer reliable. Test-time compute scaling, in which inference-time computation is deliberately expanded through chain-of-thought (CoT) prompting, best-of-n sampling, process-reward-model guided search, and Monte Carlo Tree Search methods, can produce dramatic capability improvements from a fixed set of model weights (Wei et al., 2022; Lightman et al., 2024; Snell et al., 2025). The same weights, evaluated at standard versus extended inference settings, can exhibit qualitatively different behaviour on safety-sensitive tasks (OpenAI, 2024). DeepSeek-R1 was trained with estimated total compute of approximately 3.5×10^{24} FLOPs (Epoch AI, 2026). Although this compute budget places the model well below the EU AI Act's 10^{25}

[1]Université de Lorraine, Inria, F-54000 Nancy, France. Correspondence to: Luiz Felipe Caparelli Piochi <luiz-felipe.caparelli-piochi@inria.fr>.

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

systemic-risk threshold, it achieves 97.3% on MATH-500 and 79.8% Pass@1 on AIME 2024, matching or exceeding OpenAI o1-1217 (96.4% and 79.2% respectively), a system widely characterised as frontier-class (Guo et al., 2025). Training compute and deployment capability are no longer reliably co-predictive in the way governance frameworks assumed. The efficient frontier of capability-per-training-compute has shifted dramatically, even though at the very highest training budgets the trained artifact remains the dominant determinant of absolute capability ceiling.

These developments expose three specific governance gaps at the intersection of TTC scaling and training-centric regulatory design. Prior work has begun to identify aspects of this tension. Ord argues that inference-at-deployment derails the current paradigm of capability governance through training-compute thresholds (Ord, 2025), and Sastry et al. establish the broader case for compute as a governance instrument while identifying the distinct computational signatures of training versus inference workloads (Sastry et al., 2024). The present paper extends this prior work in three ways. First, we construct a systematic tripartite taxonomy of the resulting governance failures, giving each gap a precise technical characterisation that maps directly to a distinct regulatory failure mode. Second, we analyse the reasoning opacity dimension in depth, a problem wherein hidden computation is not merely inaccessible but structurally unfaithful even where accessible, and which has not been examined in governance terms. Third, we derive technically implementable reforms anchored to specific mechanisms in existing governance instruments. We develop these gaps in Sections 3 through 5, analyse agentic compounding in Section 6, and propose reforms in Section 7. We draw primarily on the EU AI Act as the most comprehensively enacted regulatory text. We treat the now-revoked US Executive Order 14110 as an instance of the training-centric paradigm. We note that the US policy conversation has since moved away from training-compute reporting triggers, but the training-centric design pattern persists in compute-based export control thresholds (Sutter, 2025) and in the voluntary commitments endorsed at the Bletchley, Seoul, and Paris summits (UK Government, 2023; Summit, 2024; Élysée, 2025).

2. Background

2.1. Training Compute as a Capability Proxy

Contemporary governance frameworks treat training compute as a measurable proxy for the property they seek to regulate, namely deployment capability and associated risk. This choice is grounded in the predictive relationship documented by empirical scaling laws (Hoffmann et al., 2022). Training compute is attractive as a governance quantity because it is observable, auditable, and concentrated in a small number of measurable training runs, in contrast to downstream behaviours, which are diffuse and context-dependent (Bommasani et al., 2021).

The EU AI Act, whose general-purpose AI obligations entered into force in August 2025, operationalises this proxy by defining systemic risk as applying to any GPAI model whose cumulative training compute exceeds 10^{25} FLOPs, with providers at this tier facing enhanced obligations including adversarial testing and conditional auditor access to evaluations (European Parliament and Council, 2024). Executive Order 14110, before its revocation in January 2025, placed a reporting trigger at 10^{26} FLOPs (White House, 2023). Voluntary responsible scaling policies link enhanced safety-evaluation requirements to training-time milestones, whether through training-compute thresholds (Anthropic, 2023; 2024) or evaluated capability levels (OpenAI, 2025; Google DeepMind, 2025). In all cases the premise is the same. The trained artifact is the primary object of governance, and evaluating that artifact at a fixed point before deployment is treated as a sufficient safety gate. Post-market monitoring programmes such as pointed in the EU AI Act Article 72 are, however, configured around outputs and reported incidents rather than around inference-time computational configuration. The gaps we identify below persist within this monitoring architecture rather than alongside it.

2.2. Test-Time Compute Scaling

Test-time compute scaling refers to a family of techniques that improve model performance by allocating additional computation at inference time rather than at training time. The foundational technique is CoT prompting, which elicits step-by-step intermediate reasoning from a language model and has been shown to improve performance across a broad range of tasks without modifying model weights (Wei et al., 2022). More recent approaches extend this principle by using best-of-n sampling with outcome or process reward models to select among multiple candidate reasoning paths (Lightman et al., 2024; Zhang et al., 2025; Brown et al., 2024), and by applying tree-search methods such as Monte Carlo Tree Search to explore solution spaces across multiple inference calls (Snell et al., 2025; Xie et al., 2024).

Contemporary frontier models operationalise TTC scaling through dedicated reasoning modes (Xu et al., 2025). OpenAI's o1 and o3 models use an internal CoT process wherein the model generates an extended reasoning trace before producing a final response, with the trace withheld from end users (OpenAI, 2024). Anthropic's Claude 3.7 Sonnet provides an extended thinking mode in which the reasoning budget can be configured up to 128,000 tokens, with the thinking process optionally visible to API clients but not guaranteed to be accessible to third-party auditors (Anthropic, 2025). DeepSeek-R1 achieves reasoning perfor-

mance comparable to OpenAI o1 through reinforcement learning on CoT traces, with estimated total training FLOPs below the 10^{25} systemic-risk threshold, as documented in Section 3 (Guo et al., 2025; Epoch AI, 2026).

Snell et al. demonstrate that optimally scaling test-time compute can outperform simply scaling model size, as a smaller model with sufficient inference budget outperforms a model fourteen times larger under standard inference settings on several reasoning benchmarks (Snell et al., 2025). This result does not establish TTC as more capability-productive than pretraining in general, but instead, it reflects the relatively marginal gains available from inference-time scaling against the diminishing marginal returns of further pretraining at frontier scale. Because it formalises the decoupling of training compute from deployment capability, the governance significance of the result is structural rather than comparative. Two governance-relevant properties follow. First, the training-compute-to-capability relationship is no longer a function of a single observable quantity but a joint function of training and inference configuration. Second, empirical evidence suggests that capability gains from additional inference compute exhibit diminishing marginal returns at sufficient budget (Snell et al., 2025), implying that governance-relevant capability thresholds are likely to be crossed at moderate rather than extreme inference budgets and that multi-budget evaluations are therefore computationally tractable.

Two modes of TTC scaling are relevant for governance purposes. Internal TTC refers to extended CoT processes within a single inference call, including the hidden scratchpad architectures used by o1-class models and the extended thinking modes of Claude 3.7 Sonnet, where intermediate reasoning may be withheld from end users at the deployer's discretion. External TTC combines multiple inference calls through best-of-n sampling (Brown et al., 2024; Tang et al., 2025), self-consistency aggregation (Wang et al., 2023), and tree-search methods. The reasoning opacity gap (Section 5) arises most acutely from internal TTC, while the threshold decoupling and evaluation staleness gaps apply across both modes.

2.3. Related Work on Inference-Time Governance

The most directly adjacent work is Ord (Ord, 2025), which argues that inference scaling reshapes AI governance by allowing capable systems to reach frontier performance without triggering training-compute regulatory thresholds. Ord treats this as a single structural challenge to the compute-governance paradigm. Sastry et al. (Sastry et al., 2024) provide a complementary framework for compute as a governance instrument, identifying the distinct computational signatures of training and inference workloads. The present paper extends this prior work in three specific ways. First, rather than treating inference-at-deployment as a unified phenomenon, we identify three structurally distinct governance failure modes with different causal mechanisms and requiring different reforms. Threshold decoupling arises from miscalibrated compute triggers, evaluation staleness arises from static elicitation protocols, and reasoning opacity arises from architectural inaccessibility of intermediate computation. Second, we develop the reasoning opacity dimension as a governance problem in its own right. The internal deliberative processes of reasoning models are architecturally inaccessible to external auditors, and where nominally accessible via visible CoT, have been empirically shown to be unreliable accounts of the computation that produced the output (Turpin et al., 2023; Lanham et al., 2023). Third, we map each proposed reform to specific mechanisms in existing governance instruments, including the EU AI Act's GPAI Code of Practice and its August 2026 enforcement timeline, providing implementable technical requirements rather than paradigm-level critique.

Table 1 summarises how TTC scaling interacts with each of the three governance gaps we describe below.

3. The Threshold Decoupling Gap

Governance frameworks embody an implicit structural assumption that capability is determined by training compute. However, when sub-threshold-trained models are deployed with extended inference, they can achieve frontier-level capabilities outside the scope of regulatory scrutiny. This mismatch between governance design and deployment architecture creates a capability gap that does not require or depend on developer intent to circumvent regulation. Under the EU AI Act's systemic risk criterion, a developer whose training run falls below 10^{25} FLOPs is exempt from the enhanced obligations of Article 55 (European Parliament and Council, 2024). Yet a model trained at 10^{24} FLOPs, one order of magnitude below that threshold, may be deployed with a reasoning policy that allocates tens of thousands of tokens of intermediate computation per inference call, yielding a deployment whose effective capability profile is indistinguishable from that of a model trained at or above the threshold.

DeepSeek-R1 demonstrates this concretely. Trained with total compute of approximately 3.5×10^{24} FLOPs, it falls below the EU AI Act's 10^{25} systemic-risk threshold, yet achieves 97.3% on MATH-500 and 79.8% Pass@1 on AIME 2024, matching or exceeding OpenAI o1-1217 (96.4% and 79.2% respectively) (Guo et al., 2025). Published estimates of the training cost vary by roughly a factor of two depending on whether reinforcement-learning compute is fully included, but even upper-end estimates leave the model below the 10^{25} threshold while its deployed capability profile matches frontier systems above it (Epoch AI,

Table 1. Summary of the three governance gaps created by test-time compute scaling, the regulatory assumptions each gap undermines, and the corresponding proposed reforms.

Gap	Assumption violated	How TTC scaling creates the gap	Proposed reform
Threshold decoupling	Regulatory triggers calibrated to training compute approximate the deployment-time capability profile	Sub-threshold models achieve frontier performance via extended inference (e.g. DeepSeek-R1 at $\approx 3.5 \times 10^{24}$ FLOPs matching o1). Structural decoupling documented across model families.	Declare maximum per-call inference budget in conformity assessments. Extend regulatory thresholds to cover inference configuration.
Evaluation staleness	Pre-deployment evaluations at standard inference settings characterise the deployment-time capability profile	Capabilities can emerge at inference budgets above those used during evaluation, making safety margins narrower than evaluations indicate. Models may also adapt behaviour to evaluation conditions.	Conduct dangerous capability evaluations at maximum deployment inference budget using approved multi-budget evaluation harnesses as the reference standard.
Reasoning opacity	Reasoning-model deployments offer auditable visibility into intermediate computation	Hidden scratchpads are architecturally inaccessible. Visible CoT is stochastic, session-specific, and not reliably faithful to the underlying computation.	Require reasoning trace logging for high-risk applications. Treat faithful CoT generation as governance-critical technical infrastructure.

2026). Because DeepSeek-R1 is commercially deployed via an API, the proposed inference-budget declaration and multi-budget evaluation reforms apply directly. A governance framework that assigns enhanced scrutiny based solely on training compute would not capture this deployment as warranting enhanced oversight, despite its performance profile on both general reasoning and safety-sensitive tasks.

The DeepSeek-R1 example illustrates a structural pattern rather than an isolated case. The capability profile of a reasoning model under extended inference is a joint function of model weights and inference configuration, a consequence of the inference-scaling properties described in Section 2.2 rather than a property of any specific architecture. Snell et al. document this pattern across multiple model families, showing that smaller models with extended reasoning budgets consistently match or outperform larger models under standard inference settings on reasoning benchmarks (Snell et al., 2025). Brown et al. demonstrate the same phenomenon for repeated sampling approaches, where models far below frontier training compute achieve competitive pass rates when given sufficient sampling budget (Brown et al., 2024; Tang et al., 2025). The governance concern is therefore structural. Any deployment pairing a sub-threshold-trained model with extended inference compute is a potential threshold decoupling case. Developers may select low training budgets for economic reasons and later expand inference in response to user demand, with no intent to circumvent regulation, yet the governance exposure mirrors that of deliberate strategy. Governance concern focuses on dangerous capability domains where reasoning improvements cross qualitative thresholds with documented compute-to-risk amplification, including biological, chemical, and cybersecurity tasks.

A related but structurally distinct threshold-decoupling vector is distillation from frontier models. A student model trained at sub-threshold compute can inherit much of a teacher's capability profile without any inference amplification, as illustrated by DeepSeek-R1's reported distilled variants (Guo et al., 2025). Distillation differs from TTC scaling because the capability is fixed at training time rather than configurable at inference, meaning it is addressed by different governance instruments such as provenance disclosure and model-card declarations of distillation lineage. We flag it here as adjacent to the threshold gap and as further evidence that training compute alone underdetermines deployment capability, although a full treatment remains out of scope.

4. The Evaluation Staleness Gap

Pre-deployment capability evaluations conducted at fixed inference budgets cannot reliably characterise the capability profile of a model subsequently deployed with configurable, higher inference budgets. This creates an evaluation staleness gap that is structurally embedded in the current safety evaluation architecture. The core issue is that inference budget is a continuous parameter of the deployment configuration, yet many evaluation protocols continue to treat it as fixed. Frontier developers have begun to scale inference within their internal dangerous-capability evaluations, and some third-party evaluators have followed, but multi-budget evaluation is neither standardised across the field nor required by current regulatory protocols.

Evaluation programmes operated by national AI safety institutes and dangerous capability evaluations used by frontier developers are conducted under specific elicitation conditions that typically do not include the maximum inference budget available at deployment (Weidinger et al., 2023; National Institute of Standards and Technology, 2023; Anthropic, 2023; Perez et al., 2022). The evidence for inference-budget-dependent capability profiles is now direct. Snell et al. demonstrate that the same model weights, evaluated at different inference budgets, yield systematically different performance profiles across reasoning benchmarks, with smaller models at high inference budgets outperforming models fourteen times larger at standard inference settings (Snell et al., 2025). This result establishes that inference budget is an independent variable in capability evaluation, not an implementation detail. When a model is subsequently deployed with access to extended thinking or high-budget reasoning modes, evaluations conducted at standard settings may no longer accurately represent its ceiling capability on safety-sensitive tasks. A further complication is that models may sometimes detect evaluation conditions and adapt their behaviour, introducing an additional wedge between evaluation-time and deployment-time performance (Nguyen et al., 2025).

The staleness gap is most consequential where marginal capability improvements can cross qualitative governance thresholds. Dangerous capability evaluations for biological, chemical, radiological, and nuclear uplift, for cybersecurity tasks, and for autonomous replication are designed precisely to identify models whose performance exceeds thresholds that trigger enhanced safety measures or deployment restrictions (Anthropic, 2023; Shevlane et al., 2023). These evaluations are meaningful only to the extent that they reflect the elicitation ceiling under realistic deployment conditions. The o1 system card illustrates what is at stake. o1-preview was assessed as providing qualitatively greater biosecurity uplift than prior non-reasoning models, with the improvement attributed to extended reasoning capability (OpenAI, 2024). A threshold calibrated to non-reasoning model performance will misevaluate a deployment that enables extended reasoning, regardless of the precise mechanism. Direct within-weights evidence on dangerous capability tasks remains an open empirical gap that future evaluation research should address.

The staleness problem is particularly acute for task-based evaluations that characterise difficulty by time-to-completion horizons. Extended reasoning substantially expands the horizon of tasks a deployed system can complete (Kwa et al., 2025). Frameworks for evaluating AI agents on realistic autonomous tasks characterise task difficulty by the time a skilled human would require to complete the task, and while Leading AI organizations' internal evaluations have begun to scale inference for these benchmarks, public third-party evaluation suites and regulatory protocols still do not systematically require reporting the expanded capability horizon that extended reasoning provides (Kinniment et al., 2023). Addressing the staleness gap requires that safety evaluations be conducted at the maximum inference budget available in deployment, and that responsible scaling policies explicitly specify inference budget as a parameter of the evaluation protocol. Evaluations under standard and extended inference conditions should be reported separately so that safety thresholds can be calibrated to the capability profile that governance frameworks are actually seeking to regulate.

5. The Reasoning Opacity Gap

A third governance gap arises from the specific architecture of contemporary reasoning models. Opacity in machine learning is not new, as pre-reasoning models offered no intermediate computation to inspect and thus created a well-established right-to-explanation problem. The novelty of reasoning models lies not in opacity itself but in the partial, unreliable accessibility of intermediate computation, which creates a distinct governance challenge. Because reasoning-model deployments produce internal deliberation that is either hidden from auditors or nominally visible but unfaithful to the final output, we treat this specific combination as a unique governance-relevant property rather than a new phenomenon. The gap operates on two structurally distinct layers, and both must be addressed for reasoning trace access to satisfy governance requirements.

The first layer is architectural inaccessibility. In o1-class models, the CoT trace is generated internally and never surfaced to deployers, auditors, or affected parties. This is not a commercial access policy that could be waived by agreement, because the trace is not recorded in a form accessible outside the inference process. For these architectures, the reasoning process through which a consequential output is generated is wholly invisible to external review. The

governance analogy is the flight data recorder requirement in aviation. Without mandated logging infrastructure, no post-hoc audit is possible regardless of intent. The second layer arises even in architectures that make traces nominally visible, such as Claude 3.7 Sonnet's extended thinking mode. Extended thinking tokens are accessible only when the API client explicitly requests them, meaning deployers can suppress trace visibility without violating any current compliance obligation. More fundamentally, even when traces are fully accessible, empirical research demonstrates that CoT processes frequently diverge from the causal factors that actually determine model outputs, a phenomenon characterised as unfaithful explanation (Turpin et al., 2023; Lanham et al., 2023; Young, 2026). A trace that does not faithfully reflect the computation that produced the output provides limited assurance to an auditor regardless of how completely it is disclosed.

The two layers are independent governance problems. Structural inaccessibility in scratchpad architectures is sufficient on its own for governance failure, as no observability requirement can be met for a system whose reasoning cannot be accessed at all. The faithfulness limitation compounds the problem for nominally accessible traces, creating a second layer of failure even where access has been technically achieved. Regulatory frameworks that require transparency sufficient to allow auditors to assess how a high-risk AI system reached its outputs (European Parliament and Council, 2024) presuppose that the reasoning process is both accessible and reliable. For reasoning models currently in deployment, neither presupposition holds without explicit governance requirements targeting both layers.

An illustrative audit scenario. Consider a high-risk deployment of a reasoning model used for automated triage of loan applications under the EU AI Act's Annex III provisions. A rejected applicant invokes Article 86's right to explanation, and the deployer must demonstrate that the decision did not rely on prohibited grounds. Under the first-layer failure, if the model uses a hidden scratchpad architecture and no trace logging has been mandated, the deployer can produce only the post-hoc summary the model chose to surface, not the deliberation that produced it. Under the second-layer failure, where the deployment uses a visible CoT with full trace logging, the auditor has a token sequence, but Turpin et al. show that models can produce fluent justifications while silently conditioning on biased features in the prompt (Turpin et al., 2023). Logged traces nonetheless remain materially useful for cross-referencing inputs and reconstructing deliberative context.

Governance of high-risk automated systems in other engineering domains has addressed analogous access problems through mandatory logging obligations rather than pre-deployment escrow. Safety-critical software in aviation and financial systems is subject to audit obligations requiring that the reasoning behind consequential decisions be reconstructable for post-hoc review. The same principle applies here, with the added complication that CoT traces are stochastic and session-specific, making inference-time logging the appropriate implementation mechanism rather than source-code escrow. Even given the unfaithfulness findings of Turpin et al., Lanham et al., and Young (Turpin et al., 2023; Lanham et al., 2023; Young, 2026), mandatory trace logging delivers measurable near-term governance value. Research into mechanistically faithful CoT generation should therefore be treated as governance-critical technical infrastructure rather than exploratory research, analogous to the priority status accorded to adversarial robustness methods in existing governance frameworks (Reuel et al., 2024). Two training trends further threaten compliance. First, optimizing CoT tokens directly via process-based rewards degrades trace faithfulness and encourages superficial legibility. Second, latent reasoning models bypass the token stream entirely via continuous internal representations, structurally preventing trace logging. Regulatory frameworks must therefore require explicit disclosures regarding both CoT optimization objectives and the presence of hidden, non-tokenized reasoning layers.

6. Compound Risks in Agentic Deployment

All three governance gaps identified above are substantially amplified when reasoning models are deployed as agents, executing multi-step tasks with access to tools, memory, and external environments. Agentic deployments are increasingly common (Staufer et al., 2026). Frontier models are routinely integrated into pipelines that interleave reasoning steps with tool calls, code execution, web retrieval, and multi-model orchestration.

In an agentic setting, the threshold decoupling gap is amplified because inference compute accumulates across multiple steps. A single agentic task may involve dozens or hundreds of model calls, each potentially involving extended reasoning, such that the aggregate inference compute allocated to the task far exceeds the training compute of the underlying model. The capability profile of the resulting agentic system has no direct correspondence to any single regulatory category, and the governance framework's focus on the trained model artifact provides no mechanism for capturing the emergent properties of the full deployment.

The evaluation staleness gap becomes more severe in agentic settings because single-call evaluations do not capture behaviours that emerge from the composition of reasoning steps. A model that performs within acceptable bounds on a dangerous capability evaluation in isolation may exhibit qualitatively different behaviour when embedded in a loop with access to code execution and web retrieval, as the

combination of these capabilities enables tasks that neither the model nor the individual tools could accomplish alone (Weidinger et al., 2023; Lazer et al., 2026). Existing evaluation frameworks do not systematically require evaluation of agentic deployments, only of the base models from which they are built.

The reasoning opacity gap is deepened in agentic settings because the reasoning traces that inform consequential actions are distributed across multiple model calls, potentially across multiple model instances, and interleaved with outputs from external tools. Reconstructing an intelligible account of why an agent took a particular action requires access to this full distributed trace, which is not systematically captured or retained under current deployment practices. The provider-deployer distinction in current governance frameworks does not provide clear guidance for pipelines in which a reasoning model orchestrates specialised subagents, each of which contributes to an output that causes harm (Reuel et al., 2024).

A concrete compound-risk example. Consider a cybersecurity reasoning agent built on a sub-threshold base model, deployed with extended thinking enabled, tool access to a code execution sandbox, and web retrieval. Evaluated in isolation at standard inference budget, the base model scores below any dangerous capability threshold on a static CTF benchmark. Deployed as described, the same weights can iteratively enumerate attack surfaces, generate and test exploit payloads in the sandbox, and retrieve up-to-date CVE disclosures, producing aggregate behaviour that qualitatively exceeds its evaluated profile. All three gaps activate simultaneously. The threshold gap exempts the base model from enhanced scrutiny, the staleness gap means the single-call evaluation never observed the tool-augmented behaviour, and the opacity gap means the reasoning distributed across dozens of subagent calls is not retained in a form an auditor can reconstruct.

Consequently, agentic deployments require extending the three core reforms from Section 7 to the system level. Inference-budget declarations must aggregate across the entire pipeline to reflect the task-level compute envelope. Capability evaluations must assess the fully composed system, including tools and orchestration, rather than isolated base models. Finally, trace logging must span the entire distributed deliberation, including tool I/O and subagent boundaries, to preserve auditability. These extensions do not replace the core reforms; they specialize them to the environments where architectural gaps most heavily compound.

Table 2. Mapping inference-era governance gaps to proposed reforms and the technical artifacts required to verify compliance.

Gap	Proposed reform	Verifiable artifact
Threshold decoupling (§3)	Inference budget declaration (§7.1)	API specification bounding maximum per-call reasoning tokens and samples
Evaluation staleness (§4)	Multi-budget capability evaluations (§7.2)	Benchmark scores reported across a schedule of inference budgets
Reasoning opacity (§5)	Reasoning trace logging (§7.3)	Retained, auditor-accessible traces for consequential deployments
Agentic compounding (§6)	All three reforms jointly	Declared budgets, evaluations, and traces aggregated across the pipeline

7. Toward Inference-Aware Governance

The governance gaps described above are core features of the training-centric paradigm and require technically grounded responses. We propose three reforms, each specified as a concrete technical requirement, followed by a discussion of implementation challenges and international applicability. Table 2 maps each gap to its proposed response and the artifact a conformity assessment would need to inspect.

7.1. Inference Budget Declaration

Governance frameworks that use compute thresholds to trigger enhanced scrutiny should be extended to require declaration of the maximum per-call inference budget available in deployment. Conformity assessments and model registrations should require providers to declare the maximum reasoning budget configurable through their deployment infrastructure alongside training compute. Thresholds for enhanced scrutiny should then specify both training compute and maximum per-call inference budget, with conversion factors reflecting the qualitatively different roles of training and inference computation in determining capability.

This requirement is tractable for API-mediated deployments, where inference budgets are a configurable property of the provider's infrastructure and can be documented and verified through API specifications and technical audits without monitoring individual users. The EU AI Act's code of practice for GPAI providers, published in July 2025, with associated obligations in force from August 2025 and enforcement beginning August 2026, provides an immediate mechanism for requiring inference budget declaration (European Parliament and Council, 2024; European Commission AI Office, 2025).

Two operational points deserve explicit treatment. The de-

clared ceiling should be the provider-side configurable cap on reasoning tokens or sample count exposed through the deployment API, with each access tier (consumer, enterprise, government) declared separately, since governance-relevant capability is set by the highest accessible configuration. Open-weight models distributed without API mediation fall outside this mechanism by construction, so model-card disclosure should instead specify the inference configuration under which dangerous-capability evaluations were conducted.

Such declarations are useful only insofar as they can be independently verified. Self-reported ceilings carry little assurance value, so verification should rely on infrastructure-level evidence (signed API specifications, audit-grade logs of per-call reasoning-token counts, and randomised conformance probes by accredited auditors) rather than self-attestation. Accreditation should run through the EU AI Office in coordination with national AI safety institutes, with equivalent arrangements via the UK and US AI Safety Institutes in their respective jurisdictions. Where the model provider, the inference compute provider, and the deployment platform fall under a single corporate parent, governance frameworks should additionally require structural separation of audit evidence, by analogy with auditor-independence requirements in financial reporting.

7.2. Multi-Budget Capability Evaluations

Responsible scaling policies and the evaluation protocols used by AI safety institutes should require that dangerous capability evaluations be conducted at the maximum inference budget available in deployment, in addition to standard settings. Evaluation protocols should specify inference budget as a parameter alongside model version and alignment procedure. Capability profiles under standard and extended inference conditions should be reported separately so that safety thresholds can be set to the capability level available at deployment rather than under evaluation conditions.

Implementing this reform does not require resolving open problems in alignment research. It requires that evaluation organizations have the computational resources and procedures to run evaluations at higher inference budgets, and that developers provide access to extended reasoning modes for evaluation purposes as a condition of publication or deployment. The diminishing-returns range Snell et al. document is specific to 2024-era reasoning models, and the empirical limit may shift upward for current frontier systems. The qualitative point that inference-compute gains saturate well below pretraining-comparable scales is robust to this shift, so multi-budget evaluations remain achievable within existing evaluation programs (Snell et al., 2025).

7.3. Reasoning Trace Logging for Auditable Deployments

For high-risk AI applications, providers and deployers should be required to log full reasoning traces and provide accredited auditors with conditional access to these logs upon request. The requirement is analogous to logging obligations imposed on safety-critical software in aviation and financial systems, where intermediate computation states must be reconstructable for post-hoc incident investigation. Reasoning traces should be stored in a format that preserves the complete chain of thought, including intermediate steps, uncertainty expressions, and self-corrections.

Three feasibility concerns shape the scope of this obligation. Reasoning traces for extended-thinking deployments can be one to two orders of magnitude larger than the user-visible output, so retention should be scoped to high-risk regulated applications with retention periods calibrated to the audit window of the relevant regime. Traces typically incorporate user inputs and may themselves constitute personal data under applicable data-protection regimes, so access protocols should be structured around conditional auditor access under data-minimisation and necessity tests rather than open archival disclosure. Finally, trace logging delivers governance value only insofar as the logged content faithfully reflects the computation that produced the output, an assumption undermined by training-time optimisation pressure on chain-of-thought content and by architectures that move reasoning into latent representations (Section 5). Trace-logging obligations should therefore be paired with disclosure requirements covering both factors, so that auditors can identify deployments where logging is no longer informative.

7.4. Implementation Challenges and Verification Infrastructure

The reforms above carry implementation trade-offs. The inference budget declaration reform applies naturally to API-mediated deployments, where inference budgets are configurable and auditable through API specifications and technical audits. For open-weight models deployed without API mediation, no comparable mechanism exists. Governance should therefore initially scope inference budget obligations to API-mediated high-risk deployments, with model card disclosures requiring developers to declare the maximum inference configuration under which safety evaluations were conducted. Reasoning trace logging raises competing legal obligations, since traces incorporating user data in consequential decisions may constitute personal data subject to data minimisation and deletion rights under applicable data protection frameworks, and may contain commercially sensitive deliberative content. Logging obligations should therefore be scoped to high-risk regulated applica-

tions with access-control and proportionality review rather than applied uniformly.

Credible implementation of the multi-budget evaluation reform requires verification infrastructure that does not yet exist operationally. We envisage approved multi-budget evaluation harnesses accredited by a designated competent authority, in the EU context the AI Office in coordination with national AI safety institutes, and in the UK and US contexts the respective AI Safety Institutes. Accreditation should be conditioned on coverage of dangerous capability domains, parameterisation over inference budget and tool availability, and reproducibility on sealed evaluation sets. Resistance to evaluation gaming requires rotating test items between disclosed and sealed subsets, and running evaluations on developer-provided endpoints cross-audited for evaluation-aware behavioural drift (Nguyen et al., 2025). These provisions can be specified as technical annexes to the EU AI Act GPAI Code of Practice without requiring legislative change.

7.5. International Applicability

The governance gaps we identify are not unique to the EU AI Act. The UK and US AI Safety Institutes developed pre-deployment evaluation protocols that share the same training-centric architecture, conducting assessments under fixed inference conditions and assigning obligations to the trained artifact rather than the deployment configuration. With the revocation of Executive Order 14110 in January 2025, the voluntary commitments endorsed at Bletchley, Seoul, and Paris carry increased weight in shaping developer practice, and both frameworks presuppose accurate characterisation of deployed capability (UK Government, 2023; Summit, 2024; Élysée, 2025). Inference-budget declarations and multi-budget evaluation protocols impose the same technical requirements regardless of whether applied through binding regulation or voluntary agreement, and an inference-budget dimension can be added to existing compute-threshold calculations in export controls. The reforms are therefore multi-jurisdictionally generalisable rather than EU-specific.

8. Conclusion

Contemporary AI governance frameworks assume training-centric capability, yet deployed AI systems operate differently. We have identified three governance gaps created by test-time compute scaling. The threshold decoupling gap allows training-compute triggers to be circumvented through inference amplification, as shown by DeepSeek-R1's frontier performance below the EU AI Act's 10^{25} threshold. The evaluation staleness gap occurs because pre-deployment evaluations at fixed inference budgets fail to capture capabilities that emerge with extended reasoning, especially when models adapt to evaluation conditions. The reasoning opacity gap operates on two layers. Hidden scratchpad architectures are inaccessible by construction, and visible reasoning traces are not reliably faithful even where accessible. Agentic deployment compounds all three gaps in ways the current regulatory architecture was not designed to address.

The reforms we propose are technically grounded and do not require resolving fundamental open problems in AI safety research. Extending compute disclosures to include declared inference budgets, requiring multi-budget capability evaluations as the reference standard for safety assessment, and mandating reasoning trace logging for high-risk regulated applications are each achievable within existing governance instruments and voluntary commitment frameworks. The models and deployment patterns exposing the gaps described in this paper are not future systems. They are in widespread production use, and the governance frameworks intended to oversee them were designed for a technical paradigm that their deployment has already superseded.

Impact Statement

This paper argues for structural reforms to AI governance frameworks in response to technical developments that current frameworks were not designed to address. The gaps we identify are asymmetric in their consequences, creating blind spots that benefit developers who train capable models at low cost and deploy them with high inference budgets, while reducing the effectiveness of oversight mechanisms the public and policymakers rely on. Our reforms impose additional compliance obligations on developers and deployers of high-risk systems. We recognise that compliance costs are not trivial and that poorly designed obligations can create barriers to entry that advantage large, established providers. The reforms are therefore specified as targeted technical requirements with well-defined scope, primarily API-mediated high-risk deployments, rather than broad compute restrictions. Technical governance is only one component of effective AI governance, and addressing these gaps will also require engagement from legal scholars, civil society, and affected communities whose perspectives fall outside the scope of this paper.

LLM Usage Statement

The author used large language models as assistive tools during preparation of this manuscript, namely for light language polishing and reformatting of tables. The conceptual framing, the tripartite taxonomy of governance gaps, the analysis of governance instruments, the reform proposals, and the selection and interpretation of cited literature are the author's own work. No empirical results, citations, or

substantive claims generated by LLMs are included.

References

Anthropic. Responsible scaling policy. Technical report, Anthropic, 2023. URL https://www.anthropic.com/responsible-scaling-policy.

Anthropic. Responsible scaling policy version 3.0. Technical report, Anthropic, 2024. URL https://www.anthropic.com/news/responsible-scaling-policy-v3.

Anthropic. Claude 3.7 sonnet and claude code. Technical report, Anthropic, 2025. URL https://www.anthropic.com/news/claude-3-7-sonnet.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. URL https://crfm.stanford.edu/assets/report.pdf.

Brown, B., Juravsky, J., Ehrlich, R., Clark, R., Le, Q. V., Ré, C., and Mirhoseini, A. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Epoch AI. Data on AI models, March 2026. URL https://epoch.ai/data/ai-models. Accessed: 2 Apr 2026.

European Commission AI Office. General-purpose AI code of practice. Technical report, European Commission, 2025. URL https://code-of-practice.ai/.

European Parliament and Council. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act). Official Journal of the European Union, L Series, June 2024. Articles 51, 55, 13.

Google DeepMind. Frontier safety framework version 3.0. Technical report, Google DeepMind, September 2025. URL https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3.pdf.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. v. d., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030, 2022.

Kinniment, M., Sato, L. J. K., Du, H., Goodrich, B., Hasin, M., Chan, L., Miles, L. H., Lin, T. R., Wijk, H., Burget, J., et al. Evaluating language-model agents on realistic autonomous tasks. *arXiv preprint arXiv:2312.11671*, 2023.

Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Von Arx, S., et al. Measuring ai ability to complete long software tasks. *Advances in Neural Information Processing Systems*, 38: 92213–92266, 2025.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Lazer, S. J., Aryal, K., Gupta, M., and Bertino, E. A survey of agentic AI and cybersecurity: Challenges, opportunities and use-case prototypes. *arXiv preprint arXiv:2601.05293*, 2026.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

National Institute of Standards and Technology. Artificial intelligence risk management framework (ai rmf 1.0). Technical Report NIST AI 100-1, U.S. Department of Commerce, January 2023. URL https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

Nguyen, J., Hoang, K., Attubato, C. L., and Hofstätter, F. Probing and Steering Evaluation Awareness of Language Models, July 2025.

OpenAI. OpenAI o1 system card. Technical report, OpenAI, September 2024. URL https://openai.com/index/openai-o1-system-card/.

OpenAI. Preparedness framework version 2.0. Technical report, OpenAI, 2025. URL https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf.

Ord, T. Inference scaling reshapes AI governance. *arXiv preprint arXiv:2503.05705*, February 2025.

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., and Irving, G. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3419–3448, 2022.

Reuel, A., Soder, L., Bucknall, B., and Undheim, T. A. Position: Technical research and talent is needed for effective AI governance. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42543–42557. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/reuel24a.html.

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., O'Keefe, C., Hadfield, G. K., Ngo, R., Pilz, K., Gor, G., Bluemke, E., Shoker, S., Egan, J., Trager, R. F., Avin, S., Weller, A., Bengio, Y., and Coyle, D. Computing power and the governance of artificial intelligence. *arXiv preprint arXiv:2402.08797*, 2024.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., et al. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*, 2023.

Snell, C. V., Lee, J., Xu, K., and Kumar, A. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Staufer, L., Feng, K., Wei, K., Bailey, L., Duan, Y., Yang, M., Ozisik, A. P., Casper, S., and Kolt, N. The 2025 AI agent index: Documenting technical and safety features of deployed agentic ai systems. *arXiv preprint arXiv:2602.17753*, 2026.

Summit, A. S. Seoul declaration for safe innovative and inclusive ai, 2024.

Sutter, K. M. U.s. export controls and china: Advanced semiconductors. Technical Report CRS Report R48642, Congressional Research Service, Washington, D.C., sep 2025. URL https://crsreports.congress.gov/product/pdf/R/R48642. Version 6, Updated Sep. 19, 2025.

Tang, Y.-C., Chen, P.-Y., and Cavallaro, A. CarBoN: Calibrated best-of-N sampling for test-time reasoning. *arXiv preprint arXiv:2510.15674*, 2025.

Trager, R., Harack, B., Reuel, A., Carnegie, A., Heim, L., Ho, L., Kreps, S., Lall, R., Larter, O., hÉigeartaigh, S. Ó., et al. International governance of civilian ai: A jurisdictional certification approach. *arXiv preprint arXiv:2308.15514*, 2023.

Turpin, M., Michael, J., Perez, E., and Bowman, S. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ed3fea9033a80fea1376299fa7863f4a-Paper-Conference.pdf.

UK Government. The Bletchley declaration by countries attending the AI safety summit. Technical report, UK Government, November 2023. URL https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration.

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., et al. Sociotechnical safety evaluation of generative AI systems. *arXiv preprint arXiv:2310.11986*, 2023.

White House. Safe, secure, and trustworthy development and use of artificial intelligence. *Federal Register*, 88 (211):75191–75226, November 2023.

Xie, Y., Goyal, A., Zheng, W., Kan, M.-Y., Lillicrap, T. P., Kawaguchi, K., and Shieh, M. Monte Carlo Tree Search boosts reasoning via iterative preference learning. *arXiv preprint arXiv:2405.00451*, 2024.

Xu, F., Hao, Q., Shao, C., Zong, Z., Li, Y., Wang, J., Zhang, Y., Wang, J., Lan, X., Gong, J., et al. Toward large reasoning models: A survey of reinforced reasoning with large language models. *Patterns*, 6(10), 2025.

Young, R. J. Lie to me: How faithful is chain-of-thought reasoning in open-weight reasoning models? *arXiv preprint arXiv:2603.22582*, 2026.

Zhang, Z., Zheng, C., Wu, Y., Zhang, B., Lin, R., Yu, B., Liu, D., Zhou, J., and Lin, J. The lessons of developing process reward models in mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10495–10516, 2025.

Élysée. Statement on inclusive and sustainable artificial intelligence for people and the planet. Press release, AI Action Summit, February 2025. URL https://www.elysee.fr/en/emmanuel-macron/2025/02/11/statement-on-inclusive-and-sustainable-artificial-intelligence-for-people-and-the-planet.