UNEC Journal of Computer Science and Digital Technologies
Volume 2, № 1, 2026, pp. 76-99
https://doi.org/10.30546/UNECCSDT.2026.001.263

Section:
Digital Transformation and Emerging Technologies

Making Inference Infrastructure Public: A Three-Layer Production Model for Small and Mid-Sized States

Oktay Ibrahimov
TF Consultancy, Baku, Azerbaijan
0009-0004-4033-1952, ogtay.ibrahim@outlook.com

Abstract

As AI evolves from an applied technology into the foundational substrate of economic coordination and public governance, the central strategic imperative for states has shifted: from ownership of frontier models to reliable, governable access to large-scale inference capacity. This paper extends the AI as Public Infrastructure framework by introducing Inference Infrastructure (I&I)—the nationally embedded capacity to generate and operationalize machine-mediated reasoning at scale—and argues it emerges only through synchronized co-development of three interdependent layers: physical compute and data infrastructure, operational AI systems capability, and institutional demand architecture. The absence of any single layer produces systemic underperformance, formalized through the multiplicative capacity function $I = C \times S \times D$. Integrating this I&I model with the Infrastructure Status Index, the paper introduces the *asynchrony penalty* — the systematic loss of inference capacity that follows from uneven development across the C, S, D layers — as the structural explanation for the documented gap between national AI strategy ambition and observable AI integration across the developing and middle-income world. The paper offers a diagnostic and policy framework for small and mid-sized states navigating the structural tension between digital sovereignty and integration with the global AI ecosystem. Drawing on evidence from the global shift toward inference-dominant AI workloads and national AI strategies—including Azerbaijan's AI Strategy 2025–2028 and Digital Economy Development Strategy 2026–2029, we demonstrate that the strategic objective for non-frontier states is not technological prestige but institutionalized access to machine reasoning capacity under conditions of governed interdependence.

***Keywords*:** AI governance, Inference Infrastructure, Public Infrastructure, Digital Sovereignty, Inference Economy, Governed Interdependence

Received:	*Revised:*	*Accepted:*	*Published:*
29/05/2026	*03/06/2026*	*06/06/2026*	*17/06/2026*

1. Introduction: From AI Adoption to Inference Capacity

1.1 The Fragmentation of AI Policy Discourse

National AI strategies — across advanced and developing economies alike — converge on a familiar repertoire: frontier model development, research ecosystem cultivation, startup acceleration, regulatory framework construction, and talent pipeline investment. Each element contributes to AI capability in isolation. Their disaggregated pursuit produces what we term *strategic scatter*: resource allocation across AI-adjacent objectives without a specified theory of how they combine to produce systemic economic output. In its absence, mid-sized and developing states default to policy mimicry: adoption of frontier-state strategic forms (ethics frameworks, talent pipelines, startup ecosystems, regulatory bodies) without the capital base, institutional density, or absorptive capacity that give those forms functional content. A special

case of institutional isomorphism (DiMaggio and Powell, 1983), AI-policy mimicry is reinforced by the standardization of strategy templates circulated through multilateral organizations, consultancies, and development banks—producing global convergence in AI strategy form that masks systematic divergence in substance.

The persistence of mimicry despite documented failure rates—McKinsey (2025) estimates that 70–80% of agentic AI initiatives fail to scale beyond pilot, and the Federal Reserve Board (2025) documents systematic gaps between strategic commitment and realized AI integration—signals a structural constraint embedded in the current institutional environment. Correcting this requires a shift in the strategic object: from aggregated AI capability to inference capacity understood as a production output.

1.2 The Inference Turn

A structural shift in AI workload composition has reframed from the strategic question facing states. As AI mediates administrative decisions, supply chain optimization, energy management, legal drafting, financial modeling, and public service delivery, the relevant macroeconomic question is no longer whether a country develops AI but whether it can sustain domestically governable inference capacity embedded in institutional workflows.

Three terms carry analytical weight throughout. *Inference capacity* denotes the volume of machine-mediated reasoning operations a national system produces per unit time, capturing realized (operationalized) reasoning rather than latent capability, and remaining analytically distinct from model ownership and aggregate compute resources. *Domestic governability* is the subset of inference operations over which a state retains effective oversight, audit access, and continuity guarantees, resilient to vendor discretion, geopolitical realignment, or supply-chain disruption—positioned between territorial sovereignty (operations need not be domestically hosted) and mere commercial access. Workflow embedding is the structural integration of inference outputs into institutional decision processes such that their removal would degrade operational performance — the productive condition that distinguishes *inference infrastructure (I&I)* from *inference services*.

Non-frontier states, used throughout, denotes states that lack the capital scale, dataset access, and engineering depth to sustain frontier model pre-training, and whose strategic AI question is therefore deployment and governance rather than model ownership — a category that includes nearly all states outside the US and China and the conceptual focus of this paper. The category is internally heterogeneous along resource endowment, talent base, and geopolitical positioning — variation section 7 organizes through five trajectory profiles.

The empirical case for treating **inference capacity** and **domestic governability** as the relevant strategic variables rests on a documented inversion in the composition of AI compute. In 2023, model training accounted for approximately two-thirds of global AI compute, with inference comprising the remaining one-third. By 2025, this ratio had reversed, with inference projected to further expand its dominance through 2026 and beyond (Deloitte, 2025). This shift repositions the strategic locus of AI competition from capability formation (model training) to capability deployment (inference). Industry forecasts converge on inference market growth from approximately $113 billion in 2025 to over $250 billion by decade's end, with inference-optimized chip sales projected to exceed $50 billion in 2026 and AI data center capital expenditure reaching $400–450 billion globally (Deloitte, 2025). Power availability—defined as the provision of reliable, scalable, and cost-efficient electricity—has become the binding constraint, superseding raw compute capacity, as efficiency metrics shift from peak FLOPS to "tokens per watt per dollar."

The structural logic is straightforward: frontier model training is finite and periodic; inference is continuous, scaling with every API call and embedded workflow. The strategic implication is consequential: under model training dominance, competition centers on model ownership—shaped by capital scale, dataset access, and frontier engineering talent. Under inference of dominance, the binding constraint shifts to deployment, integration, and

governance capacity—competencies more evenly distributed and accessible to states that cannot sustain frontier model investment.

AI has entered its infrastructural phase not because models have grown larger but because inference outputs increasingly function as operational inputs into governance and economic coordination — a transition from AI as a tool to AI as a systemic substrate. Within the AIPI framework (Ibrahimov, 2025a), this corresponds to the movement toward public infrastructure status, requiring conceptual and measurement instruments adequate to its material substrate.

1.3 Scope and Contribution

This paper develops a framework for *inference infrastructure* (I&I) — the durable national capacity to generate, govern, and integrate machine-mediated reasoning into economic and administrative processes — and provides instruments for its diagnostic assessment. This contribution is fourfold.

First, the paper introduces I&I as a distinct analytical category, irreducible to compute, operational capability, or institutional demand, and emerging only through their synchronized development—relocating the strategic variable for non-frontier states from model ownership to operationalized inference under governable access.

Second, the paper specifies the multiplicative condition under which inference capacity is generated, captured heuristically as $I = C \times S \times D$. From it we derive the asynchrony penalty: the systematic loss of inference capacity from uneven layer development, and the structural explanation for the gap between AI strategy ambition and observable integration.

Third, the paper extends the Infrastructure Status Index (ISI) (Ibrahimov, 2025a) with inference-specific indicators that operationalize each of the three layers. These indicators are designed for diagnostic use: a state can identify which layer constitutes its binding constraint and adjust strategy accordingly. The extension converts AIPI from a conceptual framework into a measurable policy instrument.

Fourth, the paper applies the framework through illustrative comparative mapping across five small and mid-sized state (SMS) trajectories — Estonia, Singapore, the United Arab Emirates, Indonesia, and Azerbaijan. The cases extend the comparative mapping developed in companion work (Ibrahimov, 2026) by recoding each through the three-layer lens, and are selected to vary along the (C, S, D) profile. They are heuristic exemplars rather than empirical tests, consistent with the methodological orientation of the broader research program. Section 7 develops each case; the binding-constraint identification and strategic prescriptions follow in sections 9 and 10.

2. Theoretical Foundations and Literature Review

The framework developed in Section 4 sits at the intersection of four literatures, none of which individually provides an adequate account of inference as a productive variable: general purpose technology (GPT) theory, infrastructure economics, the political economy of compute, and digital sovereignty. We review each in turn, identifying both the analytical resources we draw on and the specific gap that motivates the present model. Jointly, these literatures establish that inference is consequential, that its provision exhibits public-goods characteristics, that its supply chain is structurally concentrated, and that sovereignty over it is contested — but none specifies the productive interaction among the inputs that determine whether a state can sustain it.

2.1 AI as General-Purpose Technology — and Its Analytical Limit

The classification of AI as a General-Purpose Technology (GPT) in the tradition of Bresnahan and Trajtenberg (1995) provides the standard framework for understanding its economic significance. Like electricity and digital computing before it, AI exhibits the three GPT hallmarks: pervasiveness across sectors, sustained potential for technical improvement, and the capacity to spawn complementary innovations. While the GPT framework captures the

broad economic significance of AI and places it within a familiar lineage of general-purpose technologies, it is analytically insufficient for the present purpose. By treating technology as exogenous and centering diffusion dynamics—adoption lags, complementary innovation, and sectoral adjustment—it reduces institutional embedding to a downstream outcome rather than a constitutive element of production. In doing so, it abstracts away from the mechanisms and institutional logic through which states and organization's structure, govern, and sustain deployment at scale. Consequently, the framework remains confined to stylize historical comparison and specifies neither a production function for deployment within national institutional systems nor operational instruments for diagnosing why deployment fails.

A more recent line of work begins to address these limitations. Ide and Talamas (2025) provide the closest analytical precursor to the framework developed here. They model AI in the knowledge economy by treating inference compute—measured in FLOPS or token throughput—as a critical production input whose economic effects are mediated by organizational structure. Their model yields two results we adopt: first, that compute functions as a general-purpose productive resource; and second, that its impact depends not only on its quantity but on the organizational architecture through which it is deployed.

The present paper extends this framework into two dimensions. First, institutional architecture is treated as a coequal argument rather than a mediating variable. Second, operational capability is introduced as a distinct third argument, whereas Ide and Talamas subsume it within their organizational variable. This distinction is necessary because operational capability and demand architecture exhibit systematically different lead times, mobility properties, and failure modes—differences that Section 5 shows to be policy-relevant.

2.2 Infrastructure Economics and the Goods Classification Problem

Classical infrastructure economics classifies goods along the dimensions of excludability and rivalry, distinguishing private, club, and public goods. AI inference occupies a distinctive position within this framework. Raw compute is rival and excludable and thus constitutes a private good. Trained model weights are non-rival and non-excludable, though often rendered excludable through licensing and access controls, approximating a public good. Inference outputs embedded in institutional workflows are rival in capacity at the point of delivery yet generate non-rival positive externalities through productivity improvements that diffuse across sectors. Inference and infrastructure—the layered system enabling inference at institutional scale—is therefore best understood as an impure public good with congestion effects, analogous to electrical grids and telecommunications networks.

This classification has direct analytical consequences. Impure public goods with congestion are characterized by a structural tendency toward private underinvestment. Because private actors internalize only a portion of the social value generated, capacity provision systematically falls short of the welfare-optimal level. This property underpins the standard infrastructure economics rationale for public investment, regulatory oversight, and standardized access regimes. Historical transitions in telecommunications, electricity, and digital broadband illustrate this dynamic, whereby privately developed technologies are progressively reorganized as publicly governed infrastructures (Frischmann, 2012; Star and Ruhleder, 1996).

Emerging policy practice supports this classification: the EU's AI Factories initiative treats AI compute as shared public infrastructure, and South Korea's 2025 commitment to deploy over 260,000 GPUs across sovereign clouds represents national infrastructure investment rather than procurement.

We accept the impure-public-good classification as established. The infrastructure economics literature specifies provision rules for systems whose impure-public-good character is already operative — but not the structural conditions under which a national system attains that character. A state may invest in computing and produce neither systemic externalities nor congestion-relevant capacity if the institutional and operational layers required to convert compute into systemic inference output are absent. Section 4 specifies the conditions under

which inference inputs combine to produce a system whose outputs warrant infrastructure-economics treatment in the first place.

2.3 The Political Economy of Compute: Concentration as a Production Constraint

The hardware supply chain for AI exhibits extreme concentration along its critical segments. Industry analyses consistently estimate NVIDIA's share of the AI GPU market in the range of 90–95%, with TSMC fabricating the overwhelming majority of advanced AI chips at the process nodes capable of supporting frontier inference. Three hyperscale cloud providers — Amazon Web Services, Microsoft Azure, and Google Cloud — mediate compute access for the bulk of global AI users, with a fourth tier of specialized providers (Oracle, CoreWeave, Lambda) operating within the same hardware dependency. Brookings (2026) characterizes the resulting structure as a system of concentrated choke points spanning minerals, energy, hardware, networks, infrastructure, data, models, applications, and the cross-cutting enablers of talent and governance.

For the framework specified in Section 4, this concentration yields a direct analytical implication not formalized in the existing literature. The compute argument, C, cannot be measured as a simple quantity—GPU count, aggregate FLOPS, or theoretical throughput—but must be specified as capacity conditional on access reliability. The issue is not ownership of compute, but guaranteed continuity of access. A state may rely on external providers under normal conditions; however, when access is contingent on the political and commercial discretion of a small number of foreign firms, nominal capacity overstates the level of compute that can be relied upon under disruption. We formalize this as the distinction between *gross compute* and *governable compute*: the former denotes total technical capacity available, while the latter denotes the subset over which the state retains continuity guarantees against geopolitical realignment, vendor decisions, and supply-chain disruption. Throughout the remainder of the paper, C refers to governable compute; gross compute appears only as a measurement caveat.

Luitse (2024) provides the anchor: cloud platforms exercise infrastructural power—the capacity to set conditions of possibility for AI production through control over the underlying stack. Federal Reserve Board (2025) quantifies the asymmetry: cumulative private AI investment 2013–2024 reached ~$470 billion in the US against $50 billion (EU), $28 billion (UK), and $6 billion (Japan), with the US hosting roughly four thousand data centers as of 2024. These differentials determine the discount factor converting gross to governable compute. A national strategy targeting a GPU count without specifying access governance has not specified C—only a number whose productive content depends on factors outside the state's control.

2.4 Digital Sovereignty: From Choice Dimensions to Production Constraints

Digital sovereignty discourse has moved beyond the binary between autonomy maximalism and market procurement. Recent work accepts that full-stack AI sovereignty is structurally infeasible even for the US and China (Brookings, 2026; Tony Blair Institute, 2026; World Economic Forum, 2026); the question has shifted to what kind of sovereignty, over which dimensions, under what conditions. McKinsey (2025) distinguishes territorial, operational, technological, and legal sovereignty; Ibrahimov (2026) extends this through governed interdependence and the Governance Membrane. None specifies the feasibility envelope for a state's (C, S, D) profile imposes — the gap between sections 4 and 8 address.

2.5 Synthesis: The Joint Gap

The four literatures identify components but none specifies the productive interaction among them: GPT theory treats embedding as diffusion outcome; infrastructure economics specifies provision rules without the conditions for attaining impure-public-good character; political economy documents compute concentration without formalizing it as a productive

constraint; sovereignty discourse specifies dimensions without the feasibility envelope. Sections 3–4 address these gaps jointly.

3. AI as Executable Epistemic Infrastructure

This section develops the conceptual apparatus required to specify the production function in section 4. We argue that AI inference, as it is currently being deployed within institutional systems, constitutes a distinctive class of productive input — *executable epistemic infrastructure* — whose properties cannot be captured by treating inference as either a technological capability or an information-processing service. The conceptual primitives developed here become the structural arguments of the production function: compute as the substrate of execution, operational capability as the conversion mechanism, and institutional demand as the productive locus.

3.1 The Mechanization of Inferential Labor

Successive techno-economic paradigms have mechanized progressively more cognitive labor: electrification mechanized physical labor, digital computing mechanized arithmetic labor, AI mechanizes inferential labor — the synthesis and production of structured judgments where operations cannot be specified as deterministic procedures. Bureaucracies and professional bodies have historically performed this labor through institutional hierarchies (Garicano and Rossi-Hansberg, 2015), bounded by headcount and coordination cost. AI removes this scaling limit at the technical level only; institutional and governance constraints are repositioned, not removed.

3.2 What Makes I&I Executable

Earlier epistemic infrastructures—libraries, encyclopedias, statistical archives, expert systems—provided consultative outputs: structured knowledge that human users queried, interpreted, and applied at human speeds. They provided inputs to institutional cognition; they did not perform it. AI inference, integrated into workflows, produces executable outputs—structured judgments at machine speed, in formats institutional procedures can ingest directly without human re-synthesis.

Three properties distinguish executable from consultative epistemic infrastructure: temporal coupling (decision-relevant timescales enabling operational rather than only deliberative integration); format compatibility (structured outputs—decisions, classifications, recommendations—machine systems and procedures ingest directly); and volumetric continuity (continuous generation rather than episodic consultation). These properties make inference non-substitutable once embedded: an institution reorganized around continuous, structured, low-latency inference cannot revert to consultative provision without operational degradation — the structural dependence that establishes the ISI essentiality dimension (Ibrahimov, 2025a) and triggers AIPI-grade governance obligations.

3.3 Three Conditions of Productive Inference

The mechanization of inferential labor establishes the technical possibility of executable epistemic infrastructure. It does not establish its productive realization. National systems exhibit substantial variance in whether nominal inference capability translates into productive inference output, and the variance cannot be explained by differences in technological access alone. We argue that productive inference requires joint satisfaction of three conditions, each of which corresponds to one structural argument of the production function specified in section 4.

Substrate condition. Inference requires computing capable of generating outputs at workflow-relevant temporal and volumetric scales. It is necessary but not sufficient: a state with extensive compute that cannot be governably accessed has satisfied the substrate condition only for the entities controlling that compute — the gross-vs-governable distinction (section 2.3).

Conversion conditions. Compute does not autonomously produce institutionally usable inference. Domain adaptation, fine-tuning, integration, monitoring, and pipeline maintenance require operational capability embedded in human and organizational capacity. Technical-to-institutional translation is a productive activity, not pass-through. Documented failure rates (McKinsey, 2025; Federal Reserve Board, 2025) confirm conversion as a substantial empirical challenge — failures of conversion, not compute access (section 5.2).

Demand conditions. Inference outputs not institutionally consumed do not constitute productive output. The locus is the workflow: regulations mandating AI-assisted procedures, procurement incentivizing inference-enabled delivery, administrative architectures integrating inference into decisions, and cumulative learning from sustained use. Without demand-side embedding, compute and conversion capacity produce experimental output, not infrastructure.

Each of these three conditions is necessary; none is individually sufficient. Their joint operation is what produces I&I, and the structural relationship among them — formalized in section 4 — is what determines whether a national system attains the productive scale at which infrastructure-grade analysis applies.

3.4 I&I: Inference Infrastructure as Productive Variable

The conceptual apparatus developed in this section supports the following definition.

Definition 4 (I&I). The durable national capacity to generate executable epistemic outputs at workflow-relevant temporal, format, and volumetric scales, conditional on the joint satisfaction of substrate, conversion, and demand conditions, under governance arrangements that sustain the legitimacy and continuity of institutional inference dependence.

This formalizes section 3.3's three conditions, treating governance as a continuity requirement rather than an external constraint. The conditions enter section 4 as production-function arguments; governance enters a feasibility constraint on the function range (section 8). Within the AIPI program (Ibrahimov, 2025a), I&I is the productive substrate of the transition AIPI describes — the layer at which "AI is becoming public infrastructure" acquires material content as measurable capacity.

4. The Three I&I Layers and the Multiplicative Condition

Section 3 established that productive inference depends on three jointly necessary conditions — substrate, conversion, and demand. This section translates them into the structural arguments of I&I: three layers whose joint development determines a state's productive inference capacity. We specify the multiplicative condition under which the layers interact and draw out its implications for strategic priority-setting. Following the AIPI/ISI convention (Ibrahimov, 2025a), the layers are specified at the country–sector level; national-level diagnosis aggregates over sectoral profiles that may differ substantially, a point we return in section 10.3.

4.1 The Three Layers

I&I rests on three structurally distinct layers, each corresponding to one of the productive conditions developed in section 3.

Layer I — Compute and Data Infrastructure (C). The physical substrate required for inference execution at workflow-relevant scale (see Section 3.3).

Layer II — Operational AI systems capability (S). The human and organizational capacity to convert compute into institutionally usable inference. This includes machine learning engineers, data engineers, AI systems integrators embedded within ministries, sectoral domain translators, and MLOps professionals — together with the organizational architecture that hosts them.

Layer III — Institutional demand architecture (D). The regulatory, procurement, and organizational structures create systematic demand for inference within institutional workflows. Procurement standards, sectoral mandates, regulatory clarity, data governance frameworks, public-sector training, and performance metrics together constitute this layer.

The substantive treatment of each layer — its components, strategic challenges, characteristic failure modes, and observable indicators — is developed in section 5. The remainder of this section addresses the *interaction* among the layers, which is the structural property that distinguishes I&I from any of its constituent inputs.

4.2 The Multiplicative Condition

The three layers interact multiplicatively rather than additively. We capture this interaction in the heuristic expression: **I = C × S × D,**

Where I represent inference capacity, and C, S, D represent the development levels of the three layers. The expression is a diagnostic heuristic, not an estimable production function — its purpose is to discipline strategic thinking, not support econometric inference. The multiplicative form captures the structural claim that no layer is dispensable; bounded substitutability among layers is plausible and addressed in section 4.4. It yields three policy implications: (1) all three layers are necessary; (2) balanced development outperforms concentrated investment in any single layer; (3) diagnosis of the binding constraint must precede prioritization. These are applied in section 5 and section 7.

4.3 The Asynchrony Penalty

The multiplicative condition implies a phenomenon we term the **asynchrony penalty**: the systematic loss of inference capacity that follows from uneven layer development. A state that invests heavily in one or two layers while allowing the third to lag operates substantially below the inference capacity its total investment could otherwise sustain, even if its absolute spending matches that of states with balanced layer development. The asynchrony penalty is not a hypothetical loss; it is the structural explanation for the documented gap between AI strategy ambition and observable AI integration that motivates the present paper.

The empirical signature is distinctive: states exhibiting the penalty have credible strategies — substantial commitments, announced investments, visible flagship initiatives — yet operate in workflows where AI integration remains shallow. Compute is built but underutilized; capability is trained but exported; demand architecture is announced but not enforced. Reducing the penalty requires not more investment in the visible layers but closing the gap with the lagging one.

The asynchrony penalty also clarifies the structural mechanism behind the policy mimicry problem identified in section 1.1: states that adopt frontier-state strategic templates without the institutional density to execute them replicate the form of allocation across layers without the substance of any individual layer's development. The result is a high asynchrony penalty — a structural outcome the multiplicative condition predicts, and that observed strategy outcomes confirm.

4.4 What the Multiplicative Condition Does Not Imply

Three clarifications protect against overreach. First, the multiplicative condition is a structural claim about layer-interaction direction, not a mathematical identity: capacity does not equal the product of three indices, and the expression is not econometrically estimated. Second, layers need not develop simultaneously; sequencing is consistent with the condition provided lagging layers do not remain near zero for sustained periods. Third, bounded inter-layer substitution is plausible — operational capability can partially compensate for compute scarcity — but no layer is dispensable. Once any layer is being developed, the others must be paced against it.

5. The Three Layers in Detail

This section develops each of the three layers introduced in section 4 — components, strategic challenges, and characteristic failure modes. Observable indicators are operationalized in section 6; country-specific material is deferred to section 7.

5.1 Layer I: Compute and Data Infrastructure

Components. Layer I encompass the physical substrate required for inference execution: GPU and accelerator capacity, AI-optimized data centers, edge compute, and high-bandwidth networks (compute); secure national data environments with residency and access governance, structured datasets, and interoperability standards (data); and the power generation, transmission, and load-management capacity required to sustain AI workloads at scale (energy). The three sub-components are functionally interdependent: compute without data is idle, data without compute unprocessable, both without energy inert. Layer I must satisfy the temporal-coupling, format-compatibility, and volumetric-continuity conditions (section 3.2) without which compute does not translate into infrastructural substrate.

Strategic challenges. The principal challenge is not absolute capacity but governable capacity — the subset over which the state retains continuity guarantees against geopolitical realignment, vendor decisions, and supply-chain disruption. Most small and mid-sized states will depend on imported hardware and cloud supplements indefinitely; the objective is not to eliminate dependence but to construct it under terms of preserving governance access. Energy carries particular weight: Norris et al. (2025) find the existing U.S. grid could integrate roughly 100 GW of new AI load through 2029 if workloads were scheduled with modest flexibility — implying energy and AI strategy must be co-designed. For non-frontier states, the co-design imperative is more acute, since the margin for inefficient utilization is smaller.

Characteristic failure modes. Three patterns recur. *Stranded compute*: GPU clusters and data centers built but underutilized because Layers II and III cannot convert raw compute into productive output. *Sovereign exposure*: nominal compute capacity high but governable capacity low operations technically supported but politically contingent on external actors. *Energy bottlenecks*: compute built without coordinated energy provisioning, producing utilization caps and rising marginal costs. All three reflect investment in Layer I's visible components without attention to its less visible ones.

5.2 Layer II: Operational AI Systems Capability

Components. Layer II comprises the human and organizational capacity required to convert compute and data into productive inference output. Five role-types constitute its core: machine learning engineers (model adaptation, fine-tuning, domain optimization); data engineers (pipeline management and data-quality maintenance); AI systems integrators embedded within ministries and public agencies; sectoral domain translators with training in both AI methods and a substantive domain (medicine, law, public administration); and MLOps professionals maintaining stable, auditable inference operations in production. Beyond these roles, the layer encompasses organizational capacity to develop AI applications suited to national linguistic, legal, and administrative contexts — generic frontier models trained primarily on English-language data perform unevenly on local languages, legal systems, or administrative procedures.

Strategic challenges. Operational capability cannot be purchased off the shelf or developed quickly. Lead time from investment to productive deployment exceeds Layer I (compute is installed in months; engineers and integrators require years to train and embed). International labor markets compound the challenge: jurisdictions with deeper Layer III outcompete those without talent retention. A state that invests in capability without simultaneously developing demand produces an export commodity, not a national resource — the coordination problem the multiplicative condition makes explicit.

Characteristic failure modes. *Capability of export*: operationally capable professionals migrate to jurisdictions with better deployment opportunities, leaving training costs without productive capacity. *Pilot stagnation*: AI initiatives reach demonstration stage but fail to scale; industry evidence suggests 70–80% of agentic AI initiatives struggle beyond pilot, indicating a dominant rather than marginal outcome. *Vendor capture*: external cloud providers and consultancies absorb the operational function the state failed to develop domestically, with

value flowing offshore and institutional learning lost. All three are invisible at the strategy-document level—a state can have a credible strategy, substantial compute, and visible initiatives while exhibiting all three simultaneously.

5.3 Layer III: Institutional Demand Architecture

Components. Layer III comprises the regulatory, procurement, and organizational structures that create systematic demand for inference within institutional workflows. Six components are observable: procurement standards requiring or incentivizing AI-enabled services in public contracting; sectoral programs mandating AI integration with explicit performance targets; regulatory clarity on data use that reduces the institutional risk premium on AI deployment; national data governance frameworks specifying conditions for using administrative data to train domain-specific AI; public-sector training that builds absorptive capacity; and performance metrics accounting for AI integration in evaluating institutional outcomes.

Strategic challenges. Demand for inference must be constructed; it does not arise spontaneously from compute and capability availability. Institutions exhibit predictable resistance: existing workflows are sunk-cost investments; professional identities are bound to non-AI processes; accountability structures are calibrated to non-AI outputs. Demand architecture is the set of institutional changes that make AI integration the path of least resistance. Timing is a second challenge: demand cannot precede capability (no supply to satisfy it) nor lag it (unmet capability exports). Demand and capability must be staged together.

Characteristic failure modes. *Strategy-without demand*: strategy documents, working groups, and budget commitments are announced while procurement, regulatory, and performance-metric structures remain unchanged—demand signaled but not constructed. *Sectoral isolation*: AI integration succeeds in one or two flagship sectors (often e-government) while remaining absent across the broader institutional landscape. *Regulatory paralysis*: uncertainty about data use, liability, and audit suppresses institutional appetite for AI integration even when other layers are developed, with institutions calculating the risk premium exceeds expected operational benefit and deferring indefinitely.

Assessment of the binding constraint across all three layers requires measurement instruments adequate to this structure, developed in the next section.

6. Extending the ISI: Measurement of I&I

The Infrastructure Status Index (ISI), developed within the AIPI framework (Ibrahimov, 2025a), measures the degree to which AI within a given country–sector pairing has crossed into public-infrastructure status along four dimensions: Essentiality, Embeddedness, Legitimacy, and Governance. ISI provides the diagnostic structure through which AIPI claims about infrastructural transition are converted into measurable assessments. This section extends ISI with a set of *inference-specific indicators* organized by the three-layer structure developed in section 4 and section 5. The extension allows ISI to function not only as a diagnostic of infrastructural status but as an instrument for identifying the binding-constraint layer and adjusting strategy accordingly.

6.1 The Logic of the Extension

The original ISI dimensions ask whether AI has crossed into infrastructure-grade significance. The indicators below answer the prior question: what is each layer's development level, and which is the binding constraint? A state crossing ISI threshold with a near-zero layer has reached infrastructural status under high asynchrony penalty — institutional dependence exceeding capacity to govern it. A state balanced across layers but not crossing ISI thresholds has built productive capacity without infrastructural significance; its challenge is embedding, not scaling. The indicators are designed for diagnostic use rather than cross-country ranking (section 10).

6.2 Layer I Indicators: Substrate Capacity

Three indicator concepts assess the substrate condition.

I-1: Domestic Inference Execution Ratio. The proportion of national inference workloads executed on domestically governable infrastructure (per section 2.3) — not raw compute installed but compute available under continuity-protected access conditions. *Proxies*: cloud billing data segmented by jurisdiction; data center utilization reports; government IT procurement records.

I-2: Sovereign Compute Diversification. The number and concentration of compute-supplier relationships available to domestic institutions, weighted by continuity-guarantee credibility. A state with a single hyperscale supplier scores lower than one with three diversified suppliers of comparable aggregate capacity. *Proxies*: vendor concentration ratios in public-sector procurement; multi-cloud arrangement enforceability; supply-chain audit records.

I-3: Energy-Compute Co-Adequacy. The proportion of allocated AI compute capacity sustainable under projected energy supply over a five-year horizon, accounting for grid flexibility and AI workload profiles. *Proxies*: power capacity allocated to AI workloads; grid flexibility assessments; forward energy adequacy projections.

6.3 Layer II Indicators: Conversion Capacity

Three indicator concepts assess the conversion conditions.

II-1: Inference Workforce Density. Per-capita's availability of the role-types specified in section 5.2, weighted by sectoral distribution and retention rates. Retention weighting is essential: density not retained in domestic deployment generates the capability-export failure mode. *Proxies*: labor force surveys; professional certification databases; tax-residence and employment-status data on certified professionals.

II-2: Institutional Embedding Capability. Presence and depth of AI systems integrators within ministries and major public agencies, and MLOps practice maturity in critical sectors. Capability institutionally co-located is productive in ways that capability located only in the private sector or research institutions is not. *Proxies*: organizational audits of public agencies; agency-level surveys on AI integration roles; production-system maturity assessments.

II-3: Domain Adaptation Capacity. Existence and scale of national-language model adaptation, sectoral fine-tuning, and locally produced AI applications in public-administration use — the ability to produce inference output adapted to national linguistic, legal, and administrative contexts rather than deploy generic frontier outputs as-is. *Proxies*: catalogues of nationally fine-tuned models; sectoral inventories of AI in production use; national-language NLP capacity documentation.

6.4 Layer III Indicators: Demand Architecture

Three indicator concepts assess the demand condition.

III-1: AI Procurement Penetration Rate. The share of public-sector procurement requires or incentivizing AI-enabled service delivery, weighted by contractual depth (preference, requirement, or core specification). Procurement converts institutional demand from signal into operational market. *Proxies*: procurement database analysis; contract reviews; tender specifications coded for AI requirements.

III-2: Workflow Integration Index. The degree to which AI inference outputs enter institutional decision processes as primary inputs (used directly in decisions) rather than supplementary inputs (advisory information re-synthesized by human decision-makers). Primary-input integration generates the structural dependence defining infrastructural status; supplementary integration does not. *Proxies*: agency surveys on AI use in administrative determinations; workflow analysis in priority sectors; documentation of AI outputs in regulatory and procurement decisions.

III-3: Regulatory and Governance Maturity. Comprehensiveness and enforcement strength of regulatory frameworks for AI integration — data governance, audit and transparency requirements, liability rules, and exit-rights provisions in contracts with foreign providers — the institutional structures that make AI integration viable for risk-averse public-sector institutions. *Proxies*: regulatory framework assessments; audit mechanism reviews; contract clause analysis.

6.5 Diagnostic Use

The extension supports diagnostic applications along three lines. *Identifying the binding-constraint layer*: the indicator profile reveals which of C, S, D is most depressed relative to peers or relative to the state's own development level in other layers — the prerequisite for the policy prescriptions in section 9. *Locating asynchrony*: distinct profiles (high-C/low-D, high-D/low-C, balanced) correspond to distinct strategic trajectories developed in section 7. *Tracking change over time*: repeated measurement reveals whether layer development is rebalancing toward the binding constraint or whether the asynchrony pattern is deepening.

Full operationalization — weighting schemes, sampling protocols, cross-jurisdictional comparability — is an open research program (section 10). The present contribution specifies the indicator's concepts and their structural relationship to the multiplicative condition; empirical implementation, including in-country pilots with statistical agencies, is a task for follow-up work.

7. Five Trajectories: Comparative Mapping of Layer Profiles

This section applies the three-layer model to five small and mid-sized state trajectories — Estonia, Singapore, the UAE, Indonesia, and Azerbaijan — recoded through the (C, S, D) lens to identify each case's binding-constraint layer and characteristic failure modes. The cases extend the comparative mapping developed in Ibrahimov (2026) and reuse its typology labels: Estonia as governance-led, Singapore as hub-positioned, the UAE as capital-led, Indonesia as scale-driven, and Azerbaijan as a transitional case with operational capability as the binding constraint. Azerbaijan’s profile is sketched here for comparative positioning; a detailed analysis appears in Section 10.

7.1 Estonia — Governance-Led Trajectory

Estonia represents the analytical limit of how far Layer III development can compensate for limited Layer I capacity. Its X-Road infrastructure provides one of the most institutionally embedded digital-government architectures globally, and the Kratt AI strategy (2019) extends this embedding into AI-mediated public administration. AI integration is mandated across e-government services, procurement standards specify AI-enabled service delivery, and operational redundancy with explicit service-level requirements acknowledges systemic dependency on inference outputs (Ibrahimov, 2025a). Layer II is moderately developed against population scale, with sectoral domain translators present in major regulated sectors (health, justice, finance) and mature MLOps practice in production e-government systems.

Layer I is the structural limit. Estonia does not maintain frontier-scale domestic compute and remains structurally reliant on foreign cloud infrastructure within the EU AI Factories arrangement and EU hyperscaler partnerships; governable compute is partially secured through EU-level governance but is not domestically constituted. The profile is high-D, moderate-S, low-C: productive contribution is bounded by the low C. As inference systems transition from administrative tools to cognitive infrastructure, governance strength alone may become insufficient where deeper integration with domestic compute is a productive requirement. Estonia's continued effectiveness depends on the stability of the EU compute environment — a dependence the framework recommends managing through diversified access within the EU AI Factories structure rather than treating as resolved by EU membership alone.

7.2 Singapore — Hub-Positioned State

Singapore exhibits the most balanced layer profile of the five cases and illustrates what coordinated three-layer development at small-state scale can achieve. Layer II is exceptionally developed by small-state standards, anchored by AI Singapore (AISG), the SEA-LION sovereign language model initiative, and a sustained pipeline of ML engineers and data scientists trained through university–industry partnerships. Capability retention is supported by deep institutional demand in finance, health, and public administration, with Singapore's hub role generating productive domestic deployment opportunities. Layer III is similarly mature: The Model AI Governance Framework (IMDA, 2024) institutionalizes transparency, accountability, and human-oversight requirements; procurement standards consistently incentivize AI-enabled service delivery; and AI integration has moved beyond flagship initiatives into routine operational use.

Layer I is moderate but consistent with Singapore's posture. The country does not maintain frontier-scale domestic compute and remains deeply integrated into global cloud infrastructures, but its compute relationships are diversified and contractually deep: hyperscaler arrangements are negotiated from positions of governance maturity, and operational governance over compute access exceeds what the raw footprint suggests. The profile is high-S, high-D, moderate-C, with moderate-C deliberately constructed under terms that approximate governable status. The multiplicative condition predicts strong productive output, and observed outcomes are consistent. The principal risk is that hub-positioned status depends on continued diplomatic and commercial standing in a multipolar AI landscape; the diversification logic that protects Singapore today may face stress under geopolitical bifurcation.

7.3 United Arab Emirates — Capital-Led Trajectory

The UAE provides the clearest illustration of how heavy investment in the most visible layer can produce inference capability without proportionally producing I&I. Layer I development has been aggressive and well-resourced: the 2025 US–UAE AI Acceleration Partnership granted advanced semiconductor access and a planned 5-gigawatt AI campus in exchange for explicit governance and security commitments (Ibrahimov, 2026); G42 and related sovereign-fund vehicles have built frontier-grade GPU clusters and AI-optimized data centers at a scale unusual among small states; and Falcon, the UAE-developed open-weight model series, demonstrated domestic capacity for model development. By raw compute metrics relative to population, the UAE ranks among the most-capitalized AI-infrastructure states globally.

Layer II is growing but uneven cultivated through international talent attraction and flagship initiatives (health diagnostics, energy, smart cities), with operational embedding within ministries and MLOps practice shallower than the compute footprint suggests and capability concentrated in flagship organizations. Layer III is the structurally constrained layer: integration is deepest in flagship sectors and citizen-facing services, while routine institutional demand remains thin. The UAE's earlier framing of frontier AI as a public service — free national GPT-4 access alongside domestic compute investment (Ibrahimov, 2025a) — is a creative response to the demand deficit but a substitute for institutional embedding, not its constructed equivalent.

The profile is high-C, growing-S, moderate-D. The multiplicative condition predicts diminishing returns to further compute investment relative to demand-side embedding; demand architecture, given its slower lead time, should not be deferred.

7.4 Indonesia — Scale-Driven Trajectory

Indonesia is the largest case in our comparison and the most complex layer profile. As an archipelago of more than 17,000 islands spanning several major linguistic communities, Indonesia operates under structural conditions that make uniform layer development inherently

difficult; layer development is best read as uneven across geography and sector rather than monotonically high or low.

Layer I is in an active build-out. The flagship achievement is GPU Merdeka, the national GPU infrastructure that hosts Sahabat-AI, launched as an 8B–9B sovereign SLM family by Indosat Ooredoo Hutchison and GoTo and scaled to 70B-parameter variants on open-weight foundations, supporting Bahasa Indonesia and major regional languages — Javanese, Sundanese, Balinese, Bataknese — on national compute (Indosat & GoTo, 2024; 2025). Large-scale AI-ready data centers and a proposed Sovereign AI Fund signal substantial Layer I commitment, though AI-optimized capacity remains concentrated in major urban centers. Layer II is closely coupled with Sahabat-AI's deployment, cultivated through the Indosat–GoTo partnership and public-sector training aligned with the National AI Strategy (2020) and the 2025 AI National Roadmap. The choice of open-weight foundations over frontier model development is scale-appropriate: it allocates operational capability to adaptation — fine-tuning, multilingual extension, sector-specific deployment — rather than frontier engineering. Across an archipelagic institutional landscape, adaptation capability outproduces frontier capability per dollar invested.

Layer III is uneven: Sahabat-AI powers public services, education, and regulatory compliance under data-localization and Pancasila-rooted guidelines, but demand architecture is deep in flagship initiatives and urban institutions, thinner in peripheral provinces and across judicial, regulatory, and sectoral layers. The profile is moderate-C, moderate-S, uneven-D. For scale-driven trajectories, the binding constraint is consistency across the national territory rather than level of any single layer at the national average; extending demand architecture into lagging regions yields higher returns than further compute build-out in leading regions.

7.5 Azerbaijan — Transitional Case

Azerbaijan presents a transitional profile in which compute and demand architecture are emerging while operational capability constitutes the binding constraint. Layer I is in active development, with the TransCaspian Fiber Optic project, the 2025 launch of AzInTelecom's national high-performance computing centre, and diversified cloud partnerships with AWS, Google Cloud, and Microsoft Azure (Ibrahimov, 2026). Layer III shows emerging strength: the ASAN service architecture and over 450 e-government services provide an institutional substrate for AI demand, and the AI Strategy 2025–2028 (Republic of Azerbaijan, 2025a) and Digital Economy Development Strategy 2026–2029 (Republic of Azerbaijan, 2025c) signal genuine institutional commitment. Layer II is the binding constraint: the current base of AI operations professionals — ML engineers, data engineers, integrators, MLOps specialists — remains small relative to the strategy's ambitions, and brain drain to higher-paying jurisdictions is a significant risk.

The profile is moderate-C, low-S, moderate-D. The strategic implication, developed in section 10.3, is that operational capability should be prioritized over further compute investment in the near term, with demand architecture paced against capability build-up to avoid the capability-export failure mode.

7.6 Cross-Cutting Synthesis

Two patterns emerge. The binding-constraint layer varies by context while the structural logic holds: Estonia is constrained by Layer I, the UAE by Layer III, Azerbaijan by Layer II; Singapore approaches balance and Indonesia exhibit geographic unevenness rather than a single national constraint. Visibility differentials systematically bias investment toward the most photographable layer — the UAE illustrates this; Estonia the inverse. Binding-constraint diagnosis therefore requires explicit measurement (section 6).

8. Governed Interdependence as Strategic Configuration

The multiplicative condition (section 4) and comparative mapping (section 7) establish the structural conditions under which I&I can be produced. They do not specify the strategic posture toward the global AI ecosystem within which those conditions must be met. This section closes that gap: governed interdependence (Ibrahimov, 2026) is the configuration structurally implied by the multiplicative condition for non-frontier states — not the right posture because more sovereign or more cooperative than the alternatives, but the only one consistent with the production conditions the multiplicative condition specifies.

8.1 The Structural Exclusion of the Alternatives

Two strategic postures dominate contemporary AI sovereignty discourse for small and mid-sized states. *Autonomy maximalism* aims at full domestic control over the AI value chain — chip fabrication through model training to application deployment — under the assumption that strategic security requires technological self-sufficiency. *Uncritical integration* treats AI infrastructure as a procurement decision subject to standard commercial considerations, accepting whatever supplier relationships markets generate. The multiplicative condition rules both out, on structural grounds developed below.

Autonomy maximalism is structurally infeasible: the substrate condition requires hardware supply chains concentrated at NVIDIA, TSMC, and the major hyperscale's (section 2.3). No state below frontier scale has produced inference-grade compute domestically against these dependencies, and the recent acknowledgment that full-stack sovereignty is infeasible even for the US and China (Brookings, 2026; Tony Blair Institute, 2026) confirms the limit binds at the highest levels. For small and mid-sized states, maximalism diverts resources toward the least achievable layer, deepening the asynchrony penalty rather than reducing it.

Uncritical integration is structurally vulnerable because the productive contribution of inference depends on *governable* access conditions, not on access alone. The gross-vs-governable distinction developed in section 2.3 establishes that compute available under contractual terms that can be modified by foreign vendor decisions, geopolitical realignment, or supply-chain disruption is productively discounted relative to compute available under continuity-protected access conditions. A state that procures compute on standard commercial terms without specifying governance, audit, exit-rights, or diversification provisions has not specified C in the multiplicative condition; it has specified a quantity whose productive content is contingent on factors outside its control. Where these contingencies trigger — vendor decisions to terminate access, regulatory changes in the supplier jurisdiction, supply-chain interruptions — the productive value of integration collapses, and the state's institutional dependence on inference outputs becomes operationally vulnerable. In the language of section 3.2, uncritical integration produces *inference services* but not *inference infrastructure (I&I)*.

Both alternatives fail for the same reason: neither is consistent with sections 3–4's production conditions. Maximalism mis-allocates across layers; uncritical integration mis-specifies access conditions. The structural exclusion follows from the productive conditions of inference itself, not strategic preference.

8.2 Governed Interdependence: Layer-Specific Specifications

Governed interdependence is the strategic configuration in which a state participates in the global AI ecosystem under terms that preserve governance access across each productive layer. The configuration is not a single posture, but a layer-specific set of commitments derived from the three-layer model.

At Layer I. Governed interdependence requires compute access satisfying the gross-vs-governable distinction: diversified supplier relationships (no single hyperscaler accounting for capacity the state cannot afford to lose), contractual continuity guarantees (audit access, exit rights, advance-notice provisions), domestic anchor capacity sufficient for critical-function continuity under foreign-supplier disruption, and explicit data-residency and inference-pipeline jurisdiction. The configuration is not autarkic but structurally distinct from uncritical integration: governance properties are negotiated as primary terms rather than commercial

residuals. Singapore illustrates this in mature form; Estonia in EU-mediated form; the UAE in transitional form, with quantity-vs-governability still weighted toward quantity.

At Layer II (Operational AI Systems Capability). Governed interdependence specifies that operational capability be cultivated domestically under conditions that retain it productively: workforce development calibrated to deployment rather than frontier engineering (the Indonesia open-weight adaptation approach is structurally appropriate), embedding of integrators within ministries rather than concentration in research institutions, and Layer III demand architecture sufficient to retain capability against international labor-market gravity. Layer II governance has no fully independent content — it is structurally inseparable from Layer III — because operational capability cannot be "governed" except through the demand of architecture that retains it. The capability-export failure mode (section 5.2) is therefore the structural risk against which the joint Layer II / Layer III specification must be designed.

At Layer III (Institutional Demand Architecture). Governed interdependence specifies that institutional demand be constructed under terms that preserve domestic governance over inference-mediated decisions: procurement standards that specify governance properties (audit access, explainability requirements, liability allocation) alongside performance properties; sectoral mandates with explicit accountability arrangements; regulatory frameworks that establish data governance as enabling rather than residual; and performance metrics that account for inference integration in evaluating institutional outcomes. The McKinsey (2025) four-dimensional framework discussed in section 2.4 operates at Layer III, specifying the dimensions across which sovereignty configurations vary; the present framework specifies which configurations are structurally available to states with which layer profiles.

8.3 The Governance Membrane as Operational Architecture

The strategic posture specified above requires institutional architecture adequate to its operationalization. The Governance Membrane (Ibrahimov, 2026) is developed in companion work as the reference architecture for institutionalizing governed interdependence. The present paper does not re-derive that architecture — its specification is the substantive contribution of the companion paper — but it is useful to indicate the connection between the production-function framework developed here and the governance architecture developed there.

The Governance Membrane is a layered institutional architecture — comprising oversight bodies, audit interfaces, contractual standards, and exit-rights provisions — that mediates the boundary between domestic AI systems and global AI infrastructure (Ibrahimov, 2026). It filters rather than blocks integration: data sovereignty requirements are enforced at defined interface points; inference operations affecting critical institutional functions are subject to specified domestic oversight; AI service terms are negotiated from positions of institutional agency rather than from structural dependency. The Normative Compliance Model, the Infrastructure Status Index, and the Cognitive Dependence Index operate within the membrane architecture, with ISI providing the diagnostic backbone connecting section 6's measurement framework to the operational governance architecture.

The section 4 framework and the Governance Membrane are jointly constitutive: production conditions without governance architecture produce ungovernable capacity; governance architecture without production conditions produces governance with no substrate. The task is simultaneous construction of both.

9. Policy Implications

The framework developed in sections 3–8 generates two classes of policy prescription: universal commitments that follow from the multiplicative condition and apply to all non-frontier states, and profile-specific prescriptions that follow from the binding-constraint layer identified through section 6's diagnostic apparatus and illustrated in section 7. We treat each in turn, then close with sequencing implications.

9.1 Structural Exclusions: What the Framework Rules Out

The multiplicative condition rules out two strategic postures on structural grounds, as developed in section 8.1. The corresponding policy implications are:

Avoid frontier model races. Pre-training a frontier large language model requires capital, dataset access, and engineering scale systematically inaccessible to non-frontier states — and crucially, frontier model development addresses none of the three layers that determine inference capacity: it does not provide governable compute, does not develop deployment-oriented capability (frontier engineering and deployment engineering are distinct skill sets), and does not construct institutional demand. The Indonesia approach — building atop open-weight foundations and reallocating capability to adaptation, fine-tuning, and multilingual extension — is structurally appropriate and worth generalizing.

Avoid uncritical procurement of inference services. Procurement on standard commercial terms without specifying governance properties produces inference services but not I&I (section 3.2). Procurement of inference-related services should specify the governance properties identified in section 8.2: contractual continuity guarantees, audit access, exit rights, advance-notice provisions, and explicit jurisdiction over data residency and inference-pipeline operation. The cost of negotiating these properties at contracting is substantially lower than retrofitting them after institutional dependence has developed.

9.2 Diagnose Before Investing

The most consequential universal prescription is *diagnostic primacy*: identification of the binding-constraint layer before allocation of new investment. Strategic priority cannot be assigned in absolute terms because the multiplicative condition implies that the layer most worthy of additional investment varies by national context. The diagnostic apparatus developed in section 6 — domestic inference execution ratio, sovereign compute diversification, energy-compute co-adequacy, inference workforce density, institutional embedding of capability, domain adaptation capacity, AI procurement penetration rate, workflow integration index, regulatory and governance maturity — provides the instrument through which diagnosis is conducted.

The policy implication is institutional. States that have built AI strategy infrastructure (national strategies, AI ministries, working groups) without diagnostic capacity make systematically biased investment decisions, allocating resources toward the most visible layer rather than the binding-constraint layer. The implied institutional reform is the development of AI strategy diagnostic units — cross-ministry teams with the analytical capacity to assess national layer profiles using section 6's indicators, and the institutional standing to recommend reallocation of investment based on diagnostic findings. The unit need not be large but must be analytically credible and procedurally consequential.

9.3 Design for Synchronization

The multiplicative condition implies that the three layers must be developed in coordination: investment in any one layer should be paced against the development levels of the other two to avoid the asynchrony penalty. Synchronization requires three institutional reforms.

First, AI strategy governance must span the boundaries conventionally separating compute infrastructure (digital/telecoms ministries), workforce development (education/labor), and institutional procurement (finance and sectoral ministries). Policy mimicry (section 1.1) is partly a failure of inter-ministerial coordination: each ministry pursues its remit, and the asynchrony penalty emerges from the absence of coordination across remits. Synchronization requires AI strategy bodies at the head-of-government level with mandate authority across ministries.

Second, budget cycles for the three layers must be coordinated. Compute investment cycles operate on three-to-five-year horizons (data center construction, hardware procurement).

Capability development cycles operate on five-to-ten-year horizons (workforce training, institutional embedding). Demand architecture cycles operate on shorter and longer horizons simultaneously (procurement standards can change quickly; institutional integration depth accumulates slowly). Synchronized development requires budget instruments that recognize these heterogeneous cycles and protect investments in slower-cycle layers (especially capability) from the political pressure to redirect resources toward faster-cycle visible outputs.

Third, performance metrics must reflect the multiplicative condition (the diagnostic, not estimable, structural claim of section 4.2). A state that evaluates AI strategy on compute-investment metrics alone — GPU counts, data center capacity, capital deployed — measures only one layer and rewards visibility-bias. Metrics should track inference-capacity outputs: operations executed, operational integration depth across sectors, and governance properties of the inference being conducted. The section 6 indicator system provides the metric structure; what is required is institutional commitment to use it as the primary basis for AI strategy evaluation rather than the input metrics current strategies typically employ.

9.4 Profile-Specific Prescriptions

The universal prescriptions in section 9.1 – section 9.3 apply across all non-frontier states. The framework also generates profile-specific prescriptions that vary by binding-constraint layer. We organize these by the three-layer profiles illustrated in section 7.

For Layer-I-constrained states (Estonia profile, section 7.1), the priority is governable compute access, not domestic construction. Frontier-scale data centers are structurally infeasible and unnecessary; the objective is access meeting the gross-vs-governable standard (section 2.3) — diversified suppliers, continuity guarantees, jurisdiction-protected access, and shared sovereign-compute arrangements. EU AI Factories provides one model; diversified hyperscaler arrangements under explicit governance terms; WEF's Digital Embassies for Sovereign AI concept a third. The challenge is negotiating these from positions of governance maturity rather than procurement convenience.

For Layer-II-constrained states (Azerbaijan, section 7.5), the priority is capability paced against demand architecture. Capability-export (section 5.2) is a characteristic risk. Workforce programs alone — AI academies, training, university partnerships — do not address the constraint without coupled deployment opportunities. The bundle is capability investment plus embedding mechanisms (mandatory integrator roles in ministries, MLOps career tracks in regulated sectors, public-sector deployment) plus sectoral mandates creating demand pull. Workforce and demand policy are inseparable components of a single intervention.

For Layer-III-constrained states (UAE profile, section 7.3), the priority is demanding architecture construction, with procurement reform as a primary instrument. Procurement converts demand from signal into operational market: tender specifications requiring AI-enabled delivery, performance metrics accounting for AI integration, contractual standards specifying both performance and governance properties. A state with substantial compute and growing capability that has not reformed AI procurement has built supply against demand it has not constructed; the multiplicative condition predicts diminishing returns on further compute or capability until demand catches up. Demand architecture also includes regulatory clarity on data use, sectoral mandates with performance targets, public-sector training, and metrics institutionalizing demand within routine evaluation.

For uneven profiles (Indonesia, section 7.4), the priority shifts from absolute level of any layer to consistency across the national territory. The multiplicative condition operates at the level of institutional units consuming inference — ministries, agencies, sectoral systems — and unevenness across them produces uneven returns. Instruments are sub-national capacity development, sectoral coverage beyond flagship initiatives, and institutional coverage across lagging regions. Sahabat-AI provides a structural model: locally adapted capability on domestically governable compute, scaled with attention to multilingual and sub-national contexts.

For balanced profiles (Singapore, section 7.2), the priority is interdependence governance complexity: managing multiple suppliers, jurisdictional, and sectoral integration relationships where coordination is the strategic challenge. Instruments are negotiation capacity from governance maturity, diversification against single-supplier or single-jurisdiction concentration, and analytical capacity to anticipate geopolitical fragmentation.

9.5 Sequencing

The multiplicative condition has implications for the *order* in which policy moves are made, not only their content. Two sequencing principles follow the framework.

First, diagnosis precedes investment. New strategic investments should be evaluated against the layer-profile diagnosis developed under section 9.2, not against the political salience of the proposed investment or its alignment with frontier-state strategic templates. This is the practical operationalization of the diagnostic prescription.

Second, governance specification precedes integration depth. The framework's exclusion of uncritical procurement (section 9.1) is most operationally consequential at the moment when new supplier relationships are being established. Once institutional dependence on an inference service has developed, the bargaining position of the state vis-a-vis the supplier weakens, and the governance terms of achievable post-dependence are systematically inferior to that achievable pre-dependence. The sequencing prescription is that governance properties — audit access, continuity guarantees, exit rights, jurisdictional specifications — be negotiated as primary contract terms at the inception of supplier relationships, not as retrofits after operational integration has deepened.

These principles do not specify which layer should be developed first in absolute terms (section 4.4); they specify which institutional commitments should be made first within each policy intervention: diagnose before investing; specify governance before integrating.

10. Limits, Research Agenda, and Application: The Azerbaijan Case

The framework rests on structural claims whose validity is bounded by specific conditions and whose policy implications are conditional on assumptions that bear explicit articulation. This section discusses those bounds, identifies the research program required to refine the framework further, and develops the Azerbaijan case as the worked example through which the framework's claims are most concretely applied — and through which both its diagnostic value and operational limits surface.

10.1 Boundary Conditions of the Framework

The framework's claims are bounded by three structural conditions. Each is a working assumption rather than an established fact, and each constitutes a research question whose investigation would refine the framework of policy implications.

Inference-dominant phase. The framework is specified for the period in which inference workloads dominate AI compute composition, and outputs are increasingly embedded in institutional workflows. The phase is assumed to persist through at least 2030. Technological developments restructuring composition — a return to training dominance under different architectures, or a shift toward agentic systems with qualitatively different compute profiles — would require framework revision. Policy commitments should therefore be reviewed at the cycle on which technological-phase assessments are conducted, not held permanently.

Institutional embedding. Layer III rests on the assumption that productive inference depends on institutional integration into administrative, regulatory, and organizational decision processes. If inference comes to be consumed predominantly by individual users with institutional embedding peripheral, the multiplicative condition of three-argument structure would require reconsideration. The Indonesia case (Sahabat-AI deployed across institutional and consumer use) suggests embedding remains central in current trajectories; whether this persists is open.

Cost trajectory. Per-token inference costs have declined substantially over 2023–2026. If costs decline faster than institutional integration deepens, the value of governable domestic compute relative to commodity-priced foreign inference falls; in the limit, where inference is priced as a near-zero-marginal-cost utility, the productive distinction between Layer I configurations weakens. Compute-related implications are therefore conditional on cost decline relative to integration; the posture under fast cost decline (lighter Layer I, heavier Layer III) differs from that under stable costs (balanced development), and framework-aligned states should monitor the trajectory accordingly.

10.2 The Research Agenda

The framework's contribution to policy analysis is contingent on the empirical and methodological work required to operationalize it. Four research priorities follow from the present paper.

Indicator operationalization at the country–sector level. The nine indicators in section 6 require three methodological steps: measurement protocols implementable by national statistical agencies with available data; weighting schemes that aggregate across layers consistent with the multiplicative condition without overcommitting to a functional form; and cross-country comparability standards. In-country pilots with statistical agencies in AIPI-aligned jurisdictions are the appropriate next step.

Layer interaction empirics. Whether layer interactions are best characterized as multiplicative, additive, or hybrid (with bounded but non-zero substitutability) is a substantive empirical question. Identification requires cross-country variation in layer profiles paired with measurable inference-output indicators — a data infrastructure not yet existing but implied by the operationalization work above. Stronger multiplicativity strengthens the binding-constraint logic; weaker multiplicativity admits more substitutability and weakens the universal-prescription set in section 9.

Boundary condition validation. Each boundary condition admits empirical tracking: compute composition data for the inference-dominant phase; the workflow integration index for institutional embedding; inference cost benchmarks (e.g., Epoch AI, provider disclosures) for cost trajectory. Systematic monitoring provides an early warning system for framework retirement.

Cross-paper integration. The framework operates within a broader research program — AIPI/ISI on infrastructure threshold status (Ibrahimov, 2025a), CTS on cultural prerequisites (Ibrahimov, 2025b), SINT on threshold dynamics (Ibrahimov, 2025c), the present framework on the production function, and Governed Interdependence on the institutional architecture for managing dependence (Ibrahimov, 2026). Formal integration specifying how diagnostic outputs from one inform input to another is a priority exceeding the scope of any single paper.

10.3 The Azerbaijan Case: Layer-Profile Diagnostic

Azerbaijan provides the most directly applicable test of the framework's diagnostic value: its layer profile illustrates the central diagnostic claim that the binding-constraint layer is not necessarily the most visible layer. The section 7.5 sketch positioned Azerbaijan as a transitional case with Layer II as the binding constraint; this section develops that diagnostic in greater detail, identifies sectoral variation, and surfaces the framework's operational limits in this specific context.

10.3.1 Layer I — Substantive Capacity, Strategic Underdetermination

Azerbaijan's Layer I is substantive but strategically underdetermined. The TransCaspian Fiber Optic project and Digital Silk Way position Azerbaijan as a regional digital transit hub; AzInTelecom's 2025 national HPC centre established sovereign supercomputer infrastructure; cloud partnerships with AWS, Google Cloud, and Microsoft Azure (Ibrahimov, 2026) provide diversified hyperscaler access; G-Cloud is expanding; and SABAH.city includes data center

capacity in active development—all within the Concept of Digital Development (Republic of Azerbaijan, 2025b).

Together these constitute substantive Layer I, but the governance architecture converting them into a coherent compute portfolio — explicit continuity, audit, exit-rights, and jurisdictional terms — is still in development. Per the gross-vs-governable distinction (section 2.3), Azerbaijan's gross compute is substantive while governable compute is the institutionally constituted subset. The Layer I priority is therefore governance architecture, not additional capacity. The energy endowment is a strategic advantage not yet integrated with AI compute planning at the level of section 5.1's co-adequacy logic implies.

10.3.2 Layer II — The Binding Constraint

Layer II constitutes the binding constraint in the framework's most operationally consequential sense. The AI Strategy 2025–2028 (Republic of Azerbaijan, 2025a) identifies workforce development as a central priority, providing concrete instruments — the planned AI Academy, specialized training programs, dedicated Azerbaijani NLP capacity. The Center for Analysis and Coordination of the Fourth Industrial Revolution (C4IR) provides an institutional anchor, and the existing IT workforce provides a substantive base for AI-specific capability cultivation.

Three structural challenges shape the dynamics. First, the current base of AI operations professionals is small relative to the ambitions of the AI Strategy and the demand the Digital Economy Strategy implies. Second, international labor markets create persistent brain-drain risk to higher-paying EU, Gulf, and East Asian jurisdictions, making retention rate critical to net Layer II development. Third, I&I requires role-types — sectoral domain translators, ministry-embedded integrators, MLOps specialists in regulated sectors — whose cultivation requires longer embedding cycles than generic IT workforce development. SINT (Ibrahimov, 2025c) reads Azerbaijan consistently: strong policy and infrastructural pillars alongside uneven human-capital and institutional-learning layers.

The framework's policy implication, developed in section 9.4, is that operational capability development must be paced against demand architecture construction to avoid the capability-export failure mode. The AI Academy and related instruments address the capability-supply side; their effectiveness depends on the construction of domestic deployment opportunities — through procurement reform, sectoral mandates, and institutional embedding programs — sufficient to retain capability productively.

10.3.3 Layer III — Emerging Strength, Conversion Challenge

Layer III shows emerging strength alongside a conversion challenge. The ASAN service architecture (over 450 e-government services) provides an institutional substrate; the Digital Economy Development Strategy 2026–2029 (Republic of Azerbaijan, 2025c) alongside the AI Strategy 2025–2028 (Republic of Azerbaijan, 2025a) signal genuine institutional commitment; and substantial non-oil sector investment over the past two decades indicates capacity for sustained strategic implementation across multi-year cycles.

The conversion challenge is the gap, identified in section 5.3 as the *strategy-without-demand* failure mode, between high-level strategic commitment and the procurement standards, regulatory clarity, and performance metrics that constitute *constructed* rather than *signaled* demand. The SINT analysis of Azerbaijan's sectoral configuration (Ibrahimov, 2025c) provides useful disaggregation: public administration sits in a configuration of high policy intensity but moderate societal demand — the Mandate Compliance configuration — where centralized programs were institutionalized before full cross-agency data readiness; finance and energy sit in configurations of high policy intensity matched by high societal demand — Convergent Momentum — where AI integration proceeds under coordinated supply-demand alignment. The implication for the present framework is that the binding-constraint analysis at the national level partially obscures sectoral variation: in finance and energy, Layer III is converging toward

productive configuration; in public administration, Layer III shows institutional commitment without operational depth.

The strategic priority at Layer III is therefore differentiated by the sector. In sectors where Layer III is converging (finance, energy), the binding constraint is Layer II capability sufficient to absorb available demand. In sectors where Layer III shows mandate-compliance configuration (public administration), the priority is the conversion of strategic commitment into procurement and regulatory architecture that creates operational rather than declared demand.

10.3.4 Asynchrony Pattern and Strategic Implications

The aggregate pattern is moderate-C, low-S, moderate-D, with significant sectoral variation. The multiplicative condition predicts Layer II as the national binding constraint, with investment in Layers I and III yielding diminishing returns until Layer II catches up. The section 9.4 prescription is an integrated capability-and-demand strategy: capability investment (AI Academy, training programs, NLP development) coordinated with demand architecture construction (procurement reform, sectoral mandates, embedding programs) so capability output finds productive deployment rather than exporting.

Governed interdependence (section 8.2) implies parallel Layer I commitments: supplier diversification, explicit governance specification in cloud arrangements, and energy-compute co-design enabled by Azerbaijan's energy endowment. Simultaneous engagement with multiple geocognitive power poles (Ibrahimov, 2026) — domestic compute, US and European cloud partnerships, Chinese-aligned digital corridor participation — is supported on the condition that governance architecture across these relationships is institutionally specified rather than left to bilateral emergence.

10.4 What the Azerbaijan Case Reveals About the Framework

Three observations about the framework itself follow from the Azerbaijan application.

First, the multiplicative condition's diagnostic value is sensitive to the level of analytical aggregation. At the national level, Layer II is the binding constraint; at the sectoral level, the binding constraint varies substantially. A diagnostic apparatus that operates only at the national level would mis-specify the policy's priority for finance and energy sectors, where Layer II is approaching adequacy, and Layer III demand absorption is the local constraint. The implication for the framework is that the indicator system in section 6 must be applied at the country–sector level, paralleling the AIPI/ISI architecture practice (Ibrahimov, 2025a), rather than at the country level alone.

Second, prescriptions are conditional on sustained implementation cycles whose stability is itself a strategic variable: the AI Strategy 2025–2028 and Digital Economy Development Strategy 2026–2029 depend on political-cycle continuity, sustained budget execution, and the institutional capacity-building SINT flagged (Ibrahimov, 2025c).

Third, governed interdependence (section 8) admits multiple operational pathways. Azerbaijan's simultaneous engagement across geocognitive power poles is consistent with the configuration; deeper alignment with a single pole could also be consistent under different governance architecture. The framework specifies the structural conditions; the choice among consistent configurations is strategic judgment.

11. Conclusion

The strategic question for non-frontier states in the AI era is not who built the largest models. It is who can sustain inference capacity, embed it into governance and economic processes, govern its integration responsibly, and maintain meaningful sovereignty within conditions of structural interdependence. This reframing — from AI capability as aggregate ambition to inference capacity as a specifiable productive output — is the analytical move that

distinguishes operationally substantiated strategies from those that perform the form of frontier-state ambition without the substance.

The framework specifies these conditions through three multiplicatively interacting layers — compute and data infrastructure, operational capability, and institutional demand architecture. The binding-constraint layer varies by national context and must be diagnosed before investment is allocated; the asynchrony penalty that follows from uneven layer development is the structural explanation for the gap between AI strategy ambition and observable integration. The ISI extension (section 6) provides the diagnostic apparatus.

The shift is from AI leadership as policy objective to sustained inference access as public capability — recognition that strategic value lies in the institutional capacity to convert inference into governance quality, economic productivity, and social welfare. Governed interdependence (section 8) is the posture consistent with this shift: structured engagement under terms preserving domestic governance access, rather than autonomy maximalism (structurally infeasible) or uncritical integration (producing services without infrastructure). The framework's claims are bounded by the inference-dominant phase, institutional-embedding assumption, and cost trajectory; their monitoring is itself a strategic variable. The framework specifies the production conditions of inference; what should be done depends on national circumstances it illuminates but does not resolve.

Authors' Declaration

Author contributions. Sole author; responsible for conceptualization, analysis, and writing.

Use of AI tools. ChatGPT and Claude assisted with language editing, literature-scoping (identifying sources), and document consistency. These tools were not credited as authors and did not make independent claims. All analysis, judgments, and final text were reviewed and approved by the author, who takes full responsibility for the content.

References

1. Bresnahan, T. F. and Trajtenberg, M. (1995) 'General Purpose Technologies: Engines of Growth?', *Journal of Econometrics*, 65(1), pp. 83–108.
2. Brookings Institution (2026) *Is AI Sovereignty Possible? Balancing Autonomy and Interdependence*. Washington, DC: Brookings Institution.
3. Deloitte (2025) 'Why AI's Next Phase Will Likely Demand More Computational Power, Not Less', in *Technology, Media, and Telecommunications Predictions 2026*.
4. DiMaggio, P. J. and Powell, W. W. (1983) 'The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields', *American Sociological Review*, 48(2), pp. 147–160.
5. Federal Reserve Board (2025) 'The State of AI Competition in Advanced Economies', *FEDS Notes*. Washington, DC: Federal Reserve Board.
6. Frischmann, B. M. (2012) *Infrastructure: The Social Value of Shared Resources*. New York: Oxford University Press.
7. Garicano, L. and Rossi-Hansberg, E. (2015) 'Knowledge-Based Hierarchies: Using Organizations to Understand the Economy', *Annual Review of Economics*, 7, pp. 1–30.
8. Ibrahimov, O. (2025a) 'AI as Public Infrastructure: A Critical Review of the Transition from Tool to Societal Necessity', *Current Trends in Computing*, 3(2), pp. 40–61.
9. Ibrahimov, O. (2025b) 'Cultural–Technological Synergy in the Age of AI: A Conceptual Framework for Understanding Adaptive Modernization in Transitional Societies', *UNEC Journal of Computer Science and Digital Technologies*, 1(2), pp. 5–29.
10. Ibrahimov, O. (2025c) 'Making Intelligence Public: Thresholds of Policy, Demand, and AI-Readiness', *Journal of Modern Technology and Engineering*, 10(3), pp. 189–212.

11. Ibrahimov, O. (2026) 'Digital Sovereignty in the Emerging Global Cognitive–Informational Order: Geocognitive Power Poles and Governed Interdependence', *Journal of Modern Technology and Engineering*, 11(1), pp. 61–85.
12. Ide, E. and Talamas, E. (2025) 'Artificial Intelligence in the Knowledge Economy', *Journal of Political Economy*, 133(12).
13. Indosat Ooredoo Hutchison and GoTo (2024) *Sahabat-AI: Indonesia's Sovereign AI Initiative*. Jakarta: Indosat Ooredoo Hutchison. https://www.indosatooredoo.com/en/sahabat-ai
14. Indosat Ooredoo Hutchison and GoTo (2025) *Sahabat-AI 70B Scaling Announcement*. https://www.indosatooredoo.com/en/sahabat-ai
15. Luitse, D. (2024) 'Platform Power in AI: The Evolution of Cloud Infrastructures in the Political Economy of Artificial Intelligence', *Internet Policy Review*, 13(2).
16. McKinsey & Company (2025) *The Sovereign AI Agenda: Moving from Ambition to Reality*. New York: McKinsey & Company.
17. Norris, T., Profeta, T., Patiño-Echeverri, D. and Cowie-Haskell, A. (2025) *Rethinking Load Growth: Assessing the Potential for Integration of Large Flexible Loads in US Power Systems*. Durham, NC: Nicholas Institute for Energy, Environment & Sustainability, Duke University.
18. Republic of Azerbaijan (2025a) *Artificial Intelligence Strategy of the Republic of Azerbaijan for 2025–2028*. Presidential Decree, 19 March 2025. Baku: Office of the President.
19. Republic of Azerbaijan (2025b) *Concept of Digital Development of the Republic of Azerbaijan*. Presidential Decree, January 2025. Baku: Office of the President.
20. Republic of Azerbaijan (2025c) *Digital Economy Development Strategy 2026–2029*. Baku: Ministry of Economy.
21. Infocomm Media Development Authority (2024) *Model AI Governance Framework*. Singapore: IMDA. https://www.imda.gov.sg/
22. Star, S. L. and Ruhleder, K. (1996) 'Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces', *Information Systems Research*, 7(1), pp. 111–134.
23. Tony Blair Institute for Global Change (2026) *Sovereignty in the Age of AI: Strategic Choices, Structural Dependencies and the Long Game Ahead*. London.
24. World Economic Forum (2026) *How Shared Infrastructure Can Enable Sovereign AI*. Geneva.