AI Governance Implementation Capacity: A Cross-National Analysis

Measuring Implementation Readiness Across 2,100+ Policies

Lucas Sempé

February 11, 2026

Table of contents

1 AI Governance Implementation Capacity: A Cross-National Analysis

Measuring Implementation Readiness Across 2,100+ Policies

Preface

This project began with a fundamental question about the global proliferation of artificial intelligence policies: as governments worldwide develop AI governance frameworks at an unprecedented rate, do they possess the institutional capacity to implement them effectively?

To address this question, I assessed over 2,100 policies across more than 70 jurisdictions, evaluating five critical dimensions of implementation capacity: clarity, resources, authority, accountability, and coherence. The scoring methodology employed a three-model ensemble of large language models (Claude Sonnet 4, GPT-4o, and Gemini Flash 2.0), achieving adequate inter-rater reliability (ICC = 0.83).

The findings challenge conventional assumptions about global governance capacity. While high-income countries demonstrate marginally higher capacity scores overall (Cohen's d = 0.30), this advantage disappears almost entirely when the analysis is restricted to well-documented policies (Cohen's d = 0.04). What initially appeared to be a governance capacity divide reveals itself primarily as a documentation gap. Moreover, 98% of the observed variation in capacity scores occurs within income groups rather than between them. Countries such as Brazil, Kenya, and Rwanda substantially exceed predictions based on their GDP levels, suggesting that economic development alone does not determine governance capacity. Furthermore, the analysis of policy diffusion patterns indicates that countries predominantly learn from peers at similar income levels rather than following a top-down diffusion model from wealthy nations.

This report presents the empirical evidence supporting these conclusions, along with extensive robustness checks conducted to validate the findings.

Citation: Sempé, L. (2026). *AI Governance Implementation Capacity: A Cross-National Analysis.* International Initiative for Impact Evaluation (3ie).

Data and Code: github.com/lsempe77/ai-governance-capacity

2 Introduction

2.1 The Capacity Challenge in AI Governance

2.1.1 The Implementation Gap

Between 2017 and 2025, governments worldwide produced over 2,200 AI policy initiatives catalogued in the OECD.AI Policy Observatory. While impressive in quantity, this proliferation sidesteps a question that governance research has largely neglected:

Can governments actually implement their AI policies?

Implementation capacity, defined as the institutional infrastructure enabling policy execution, remains unmeasured in AI governance. While existing research documents policy content, the presence of clarity, resources, authority, accountability structures, and coordination mechanisms necessary to translate commitments into action remains unexamined.

2.1.2 What We Measure

This study assesses each policy along five capacity dimensions:

1. **C1 Clarity**: How specific are the objectives and implementation plans?
2. **C2 Resources**: Are budgets, staffing, and technical infrastructure identified?
3. **C3 Authority**: Does the policy carry legal mandates and enforcement powers?
4. **C4 Accountability**: Are there oversight mechanisms, reporting requirements, review processes?
5. **C5 Coherence**: Is there inter-agency coordination and consistency with existing frameworks?

Scores range from 0 to 4, anchored in detailed rubrics drawn from implementation science (Mazmanian and Sabatier 1983; Pressman and Wildavsky 1973).

Most AI policies lack robust implementation infrastructure. The average capacity score across 2,100+ policies is 0.83 out of 4.0, well below the scale midpoint. However, the distributional patterns prove more revealing than aggregate statistics. The modest income-group gap effectively vanishes when the sample is restricted to well-documented policies, suggesting that apparent capacity differences may primarily reflect variation in documentation practices rather than genuine governance gaps. Within-group variation accounts for nearly all total inequality, dwarfing the between-group gap. Several developing countries—including Brazil, Kenya, and Rwanda—consistently exceed predictions based on GDP levels, while policy diffusion operates horizontally across income peers rather than through vertical transfer from wealthy nations.

The book proceeds as follows: Section 3.1 reviews the theoretical foundations; Section 4.1 outlines data and scoring (with full details in the companion *Data, Methods, and Technical Appendices* volume); Section 5.1 maps the global landscape; Section 6.1 examines determinants; Section 7.1 analyses inequality patterns; Section 8.1 traces temporal trends and diffusion; Section 9.1 presents robustness checks; and Section 10.1 and Section 11.1 discuss implications.

3 Literature Review

3.1 Theoretical Foundations

3.1.1 Implementation Science and Policy Capacity

The conceptual foundation of this study draws from implementation science, beginning with Pressman and Wildavsky's (1973) observation that well-designed programs routinely fail in execution. The field addresses a fundamental question: what features distinguish policies that achieve implementation from those that remain aspirational?

Despite decades of implementation research, a puzzling gap persists: while scholars have identified multiple conditions for effective implementation, systematic measurement of these conditions across large policy samples remains rare. Most studies examine a handful of cases qualitatively, leaving uncertain whether identified implementation barriers are generalizable or context-specific. This study addresses that gap by operationalizing implementation capacity dimensions for systematic measurement across 2,100+ AI policies.

Mazmanian and Sabatier (1983) identified six conditions for effective implementation: clear policy objectives, adequate causal theory, legal structuring, committed implementing officials, organized interest group support, and stable conditions. These conditions inform the capacity dimensions employed in this study: Clarity corresponds to their objectives condition, Resources captures their commitment requirements, Authority aligns with legal structuring, Accountability with monitoring, and Coherence with inter-agency coordination. However, the framework was developed for traditional policy domains where implementation processes are relatively stable. AI governance poses distinct challenges: rapid technological change makes "stable conditions" virtually impossible; private sector expertise concentration complicates "committed implementing officials"; and cross-border AI deployment undermines single-jurisdiction legal structuring.

Sabatier (1986) synthesized top-down and bottom-up perspectives, arguing that both formal structure and implementing strategies matter. This framework motivates the focus on policy architecture rather than policy content alone. Lipsky (1980) shifted focus to "street-level bureaucrats," the frontline workers whose discretionary decisions effectively *are* policy. For AI governance, this insight has concrete implications: even comprehensive legislation may fail if regulators lack expertise or mandate. Data protection authorities interpreting GDPR's algorithmic accountability provisions, or competition regulators assessing AI market power with limited technical staff, exemplify this challenge. The **Accountability (C4)** dimension captures whether policies constrain such discretion through monitoring and evaluation frameworks.

Grindle (1996) identified four capacity types relevant to AI governance:

Table 3.1: State capacity mapping

Capacity Type	Our Dimension	Indicators
Technical	C2 Resources	Expertise, training, technology
Administrative	C3 Authority	Legal mandate, organizational structure
Political	C5 Coherence	Cross-ministry coordination
Fiscal	C2 Resources	Budget allocation

These typologies, however, assume capacity types are empirically distinguishable. Grindle's framework treats technical, administrative, political, and fiscal capacity as separate constructs, yet AI governance may require their simultaneous deployment: technical expertise without legal authority achieves little, while legal mandate without fiscal resources remains hollow. The correlation structure among capacity dimensions thus becomes an empirical question—one this study addresses by scoring each dimension independently and examining their covariation patterns.

Fukuyama (2013) argued that governance quality is conceptually distinct from democracy or GDP, proposing measurement through government outputs. This study adopts a similar approach: measuring institutional readiness through policy quality rather than inputs such as national wealth. However, Fukuyama's output-based measurement confronts a circularity problem: governance quality affects policy outputs, but poor outputs may reflect implementation failures rather than design weaknesses. This study addresses the circularity by measuring *ex ante* institutional arrangements (designated agencies, allocated budgets, articulated procedures) rather than *ex post* implementation success.

Andrews, Pritchett, and Woolcock (2017) introduced "building state capability" through iterative adaptation, warning against imposing "best practice" from high-income countries. The empirical findings support this perspective: developing countries that score well on capacity do so through different pathways than wealthy nations. Andrews' critique, however, raises measurement challenges: if capacity pathways are context-dependent, can universal scoring rubrics meaningfully compare across jurisdictions? This study navigates this tension by scoring *presence* of capacity features (budget allocation, institutional designation, monitoring systems) without prescribing *how* those features should be structured.

3.1.2 The Capacity Gap in Digital Governance

Recent research documents persistent implementation gaps in digital regulation. Yeung (2018) shows algorithmic regulation demands technical expertise most governments lack. Katzenbach and Ulbricht (2019) demonstrates that platform governance creates enforcement challenges traditional regulators struggle to address. Yet these studies examine high-income jurisdictions (primarily EU member states and the United States), leaving unclear whether capacity deficits are universal or concentrated in resource-constrained settings. Moreover, they diagnose capacity gaps qualitatively without measuring their magnitude or comparing across policy instruments.

Dafoe (2018) argues that AI governance faces unique capacity challenges: rapid technological change outpacing regulatory adaptation, concentrated expertise in private sector rather than government, and international coordination problems where no single jurisdiction can effectively regulate global AI systems. This analysis is conceptually compelling but empirically underspecified: *which* governments lack capacity, *how much* capacity they lack, and *whether* capacity gaps are widening or narrowing remain unanswered. The argument also risks technological determinism—assuming AI's complexity inherently exceeds governmental capacity—without testing whether institutional design choices might compensate for technical complexity.

3.1.3 Measurement Challenges and Contribution

Despite theoretical progress, **systematic capacity measurement remains rare**. Existing studies rely on qualitative assessments (Cihon, Maas, and Kemp 2021) or expert surveys (Dafoe et al. 2020) that cover a handful of cases at most. This methodological limitation is not incidental: capacity assessment requires deep engagement with policy text—identifying institutional designations, interpreting budget commitments, tracing coordination mechanisms—which historically required human expertise and thus did not scale beyond small samples.

The resulting gap between theory and evidence is substantial. Implementation science offers sophisticated frameworks for understanding capacity requirements, yet lacks empirical evidence on capacity distributions across jurisdictions, policy types, or time periods. AI governance scholarship diagnoses capacity deficits but cannot quantify their severity or variation. Development economics measures state capacity through proxies (tax collection, bureaucratic quality indices) that may not capture domain-specific governance dimensions.

The LLM-based scoring approach developed in this study attempts to bridge that gap, enabling assessment across 2,100+ policies without sacrificing interpretive depth. However, automated scoring introduces validity concerns: can language models reliably interpret nuanced institutional arrangements? The inter-rater reliability analysis (ICC = 0.827) addresses this concern empirically, demonstrating that LLM ensemble scoring achieves agreement levels comparable to human expert coding.

Contribution. This study addresses three gaps in the literature. **First**, while implementation science has identified capacity conditions conceptually, systematic measurement across large policy samples remains absent. This study operationalizes five capacity dimensions and applies them to 2,100+ policies, enabling distributional analysis impossible with qualitative methods. **Second**, AI governance scholarship diagnoses capacity deficits but lacks empirical evidence on their magnitude, variation, or determinants. This study quantifies capacity levels across 70+ jurisdictions, tests whether national wealth predicts capacity, and examines whether capacity varies by policy type or jurisdiction income. **Third**, state capacity measurement typically relies on generic proxies (bureaucratic quality, regulatory quality indices) that may not capture domain-specific governance requirements. This study develops AI-governance-specific capacity indicators grounded in implementation science theory.

The findings challenge conventional wisdom: national wealth shows minimal association with implementation capacity once text quality is controlled. This contradicts the assumption, implicit in

much AI governance discourse, that developing countries systematically lag in governance sophistication. As the following sections demonstrate, capacity deficits are nearly universal—high-income countries perform only marginally better than developing countries, and both groups exhibit similar within-group variation.

4 Data and Methods

4.1 Shared Methodology

This study analyses 2,216 AI policies from the OECD.AI Policy Observatory, scored by a three-model LLM ensemble (Claude Sonnet 4, GPT-4o, Gemini Flash 2.0) on 10 governance dimensions. The full methodological details — corpus construction, scoring rubrics, inter-rater reliability, and technical validation — are documented in the companion volume:

Book 4: Data, Methods, and Technical Appendices

Key parameters for reference:

Table 4.1: Methodology summary

Parameter	Value
Corpus size	2,216 policies, 70+ jurisdictions, 2017–2025
Document retrieval	94% coverage (2,085 full texts)
Analysis-ready text	1,754 documents (79.2%), 11.4 million words
Scoring models	Claude Sonnet 4, GPT-4o, Gemini Flash 2.0
Inter-rater reliability	ICC(2,1) = 0.827 ("Excellent")
Score agreement	95.4% of scores within 1 point across models

4.1.1 Capacity Scoring Framework

Each policy was scored 0–4 on five **implementation capacity** dimensions grounded in implementation science (Mazmanian and Sabatier 1983; Lipsky 1980; Grindle 1996; Fukuyama 2013):

Table 4.2: Capacity scoring dimensions

Code	Dimension	What It Measures
C1	Clarity & Specificity	Clear objectives, measurable targets, defined scope
C2	Resources & Budget	Dedicated funding, staffing, infrastructure
C3	Authority & Enforcement	Legal mandate, penalties, compliance mechanisms

Code	Dimension	What It Measures
C4	Accountability & M&E	Reporting, evaluation, oversight bodies
C5	Coherence & Coordination	Cross-agency alignment, international coordination

Alongside the five capacity dimensions, each policy was also scored on five parallel **ethics dimensions** (E1 Ethical Framework Depth, E2 Rights Protection, E3 Governance Mechanisms, E4 Operationalisation, E5 Inclusion & Participation). The ethics scores inform cross-domain comparisons within this report and are analysed in depth in Book 2. Composite scores are unweighted means: *Capacity* = mean(C1–C5), *Ethics* = mean(E1–E5). The full scoring rubric, validation protocol, and robustness checks appear in Book 4.

Code, data, and methods: https://github.com/lsempe77/ai-governance-capacity

5 Capacity Landscape

5.1 The Global Landscape of AI Governance Capacity

With the scoring methodology validated and the corpus assembled, this section turns to the central empirical question: what does the global landscape of AI governance capacity actually look like? The analysis proceeds from aggregate distributions through income-group comparisons to regional and country-level patterns, revealing a governance landscape characterised by pervasive weakness, substantial heterogeneity, and a surprising absence of the North–South divide that conventional wisdom would predict.

5.1.1 Overall Score Distribution

The aggregate picture reveals the baseline state of AI governance capacity across all 2,216 policies.

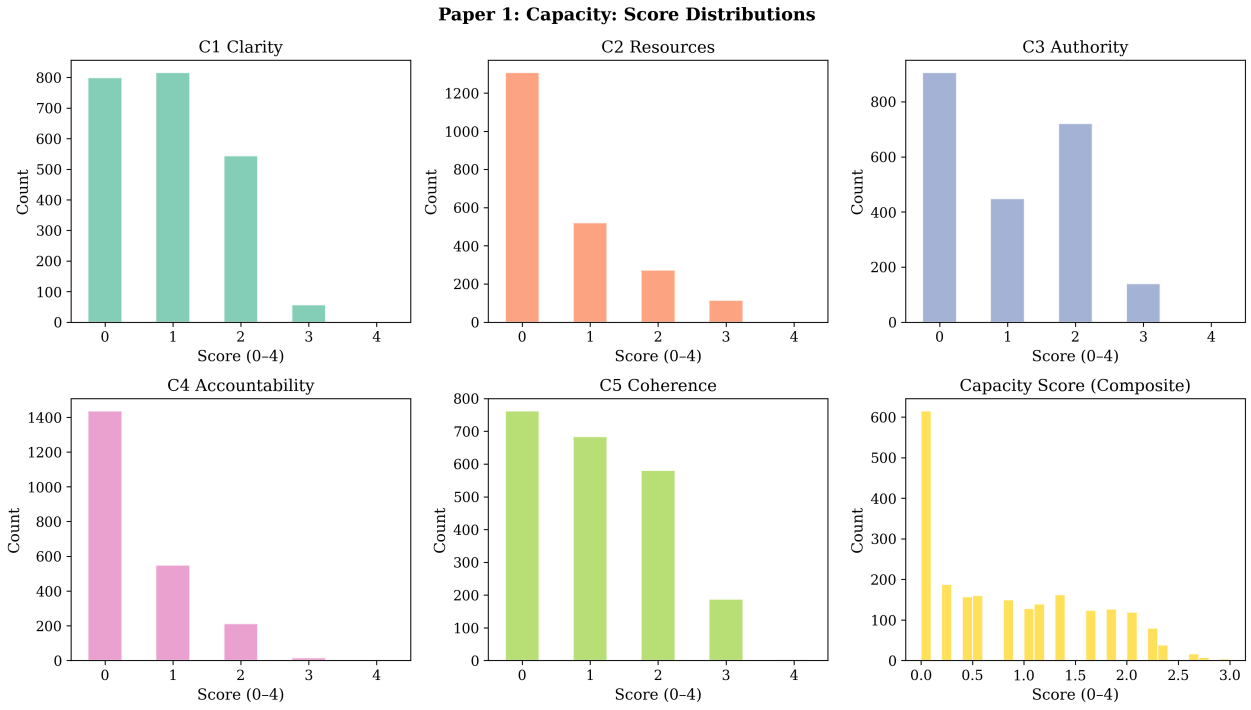

Figure 5.1: Distribution of capacity dimension scores across 2,216 policies. All five dimensions exhibit strong right skew with floor effects at zero.

The capacity composite averages **0.83/4.00** (SD = 0.77), well below the scale midpoint. However, this mean conceals substantial variation. As Figure 5.1 shows, all five dimensions exhibit right skewness, with most policies clustered at or near zero and a smaller tail extending upward. Dimension-level means range from 0.48 (C4 Accountability) to 1.07 (C5 Coherence):

Table 5.1: Capacity dimension descriptive statistics

Dimension	Mean	SD	Median
C1 Clarity & Specificity	0.94	0.97	1.00
C2 Resources & Budget	0.68	0.89	0.00
C3 Authority & Enforcement	1.04	1.08	1.00
C4 Accountability & M&E	0.48	0.72	0.00
C5 Coherence & Coordination	1.07	0.97	1.00
Capacity composite	**0.83**	**0.77**	**0.60**

One dimension stands out: **Accountability (C4) is the weakest globally**, averaging just 0.48, less than half of Coherence (C5) at 1.07. Governments are more than twice as likely to specify coordination mechanisms as to establish monitoring and evaluation frameworks. They articulate what they intend to do (Clarity) and how agencies should work together (Coherence) before committing to transparent oversight.

The standard deviations (0.72–1.08) are large, and **27.6% of policies score exactly zero** on the composite. More than a quarter of documents in the Observatory function as announcements rather than operational governance instruments. This floor effect, visible in the median of 0.00 for Resources and Accountability, motivates the Tobit approach in Section 6.1.

5.1.2 Income-Group Comparisons

The standard narrative posits a clear North–South divide: wealthy countries have sophisticated frameworks, developing countries struggle to keep up. The data complicate this story.

Figure 5.2 shows both the gap and its limits. High-income countries average 0.87 (SD 0.77), developing countries 0.65 (SD 0.72). The difference is statistically significant:

Table 5.2: Income-group capacity comparison

Metric	Value
HI mean (N = 1,700)	0.87
Developing mean (N = 397)	0.65
Welch's t	5.47
p-value	$< .001$
Cohen's d	**0.30**
Mann-Whitney U	395,388

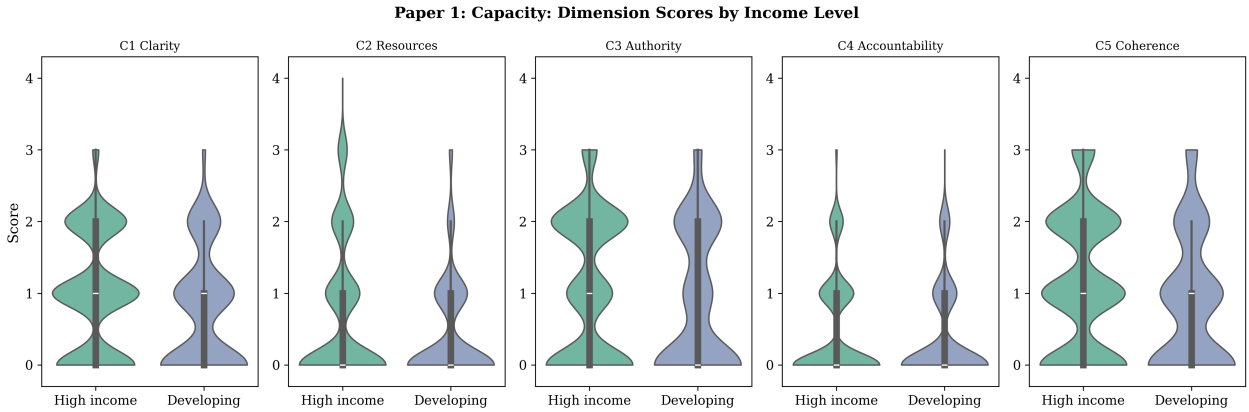

Figure 5.2: Violin plots comparing capacity score distributions between high-income and developing countries. The overlap between distributions is substantial.

Statistically significant, yes ($t = 5.47$, $p < .001$). But the Cohen's d of **0.30** is small by conventional standards, and the distributions overlap heavily. Many developing-country policies outscore the high-income median; many high-income policies sit near zero.

However, the critical finding emerges when examining text quality: this gap **vanishes entirely** for well-documented policies ($d = 0.04$, n.s.; see Section 9.1). This pattern raises a fundamental methodological question: does the observed gap reflect genuine capacity differences, or systematic variation in documentation practices across income groups?

Several mechanisms could produce weaker documentation without weaker governance. Developing countries may publish policies in national languages not captured by the OECD Observatory's predominantly English-language indexing. Their policy documents may be hosted on government websites with limited archival practices, leading to higher rates of link decay and retrieval failure. Implementation details may reside in subsidiary regulations, ministerial orders, or administrative guidance rather than in the primary strategy document indexed by the OECD. And the OECD Observatory itself may provide thinner descriptions for developing-country policies due to reporting asymmetries. These mechanisms would depress measured scores without reflecting genuine capacity deficits.

This question motivates the analyses in subsequent sections.

Dimension-level gaps. The aggregate gap masks dimension-level heterogeneity. If the divide reflected generalized institutional weakness, we would expect uniform gaps across all five dimensions. Instead, the pattern is differentiated:

Figure 5.3 visualizes how the income gap varies across dimensions, with the gap most pronounced for Resources and Coherence, and smallest for Accountability. The dimension-level statistics reveal this pattern precisely:

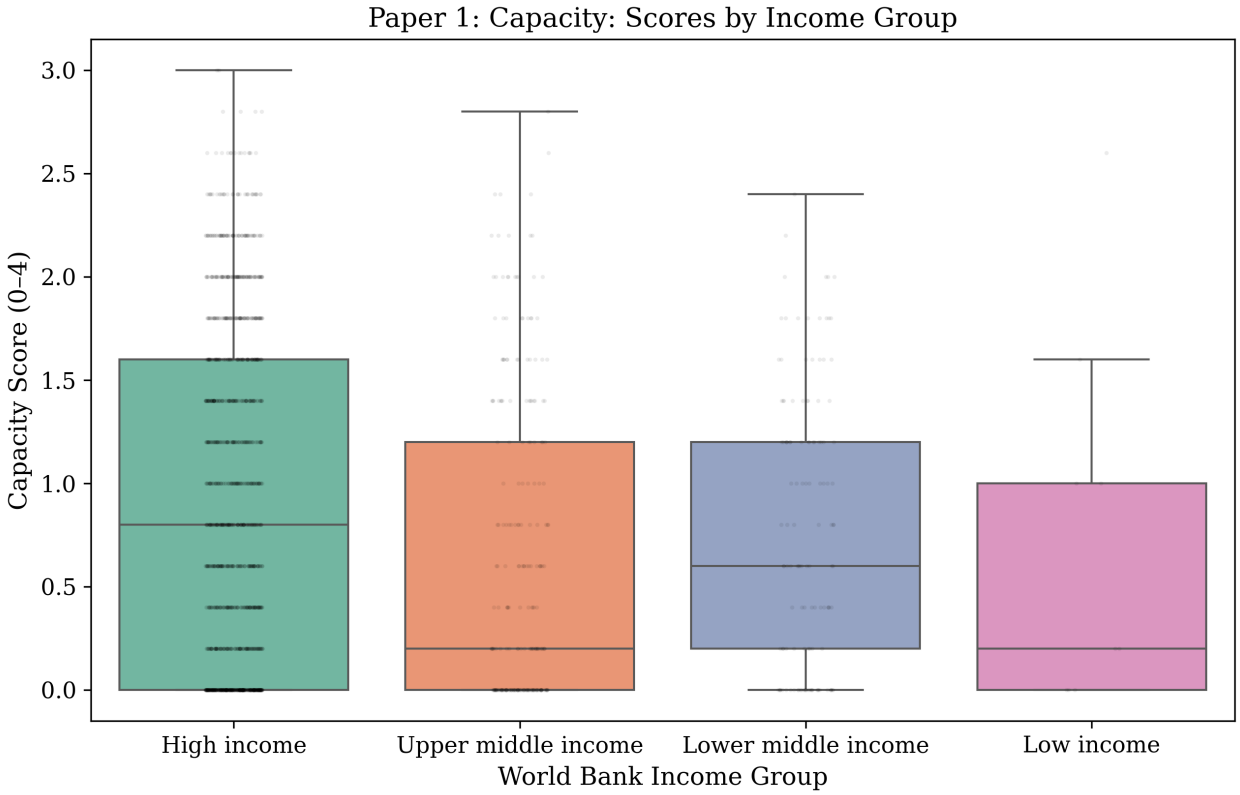

Figure 5.3: Boxplots of capacity scores by income group across all five dimensions.

Table 5.3: Dimension-level income gaps

Dimension	HI Mean	Dev Mean	Diff	d	p
C1 Clarity	0.98	0.74	0.24	0.30	$< .001$
C2 Resources	0.70	0.43	0.27	**0.32**	$< .001$
C3 Authority	1.09	0.86	0.23	0.23	$< .001$
C4 Accountability	0.48	0.37	0.10	**0.15**	.005
C5 Coherence	1.13	0.86	0.27	0.29	$< .001$

The gap is largest in **Resources (C2)** (d = 0.32) and smallest in **Accountability (C4)** (d = 0.15), confirming hypothesis H3. The logic tracks: specifying budgets and staffing requires fiscal resources that correlate with wealth. Designing monitoring frameworks, by contrast, is primarily a policy design choice, and developing countries could establish reporting requirements and oversight bodies with minimal fiscal commitment.

There is an irony here: both income groups underperform on accountability (0.48 for HI, 0.37 for developing). The reluctance to create transparent oversight mechanisms apparently transcends wealth. Accountability frameworks create political risks by enabling external assessment of implementation failures, a concern that affects all governments.

5.1.3 Regional Patterns

Income groups are blunt instruments. "Developing" lumps together Latin American countries with sophisticated regulatory traditions, South Asian nations with large tech sectors but limited governance infrastructure, and Sub-Saharan African countries with nascent AI policy ecosystems. Regional analysis provides finer grain.

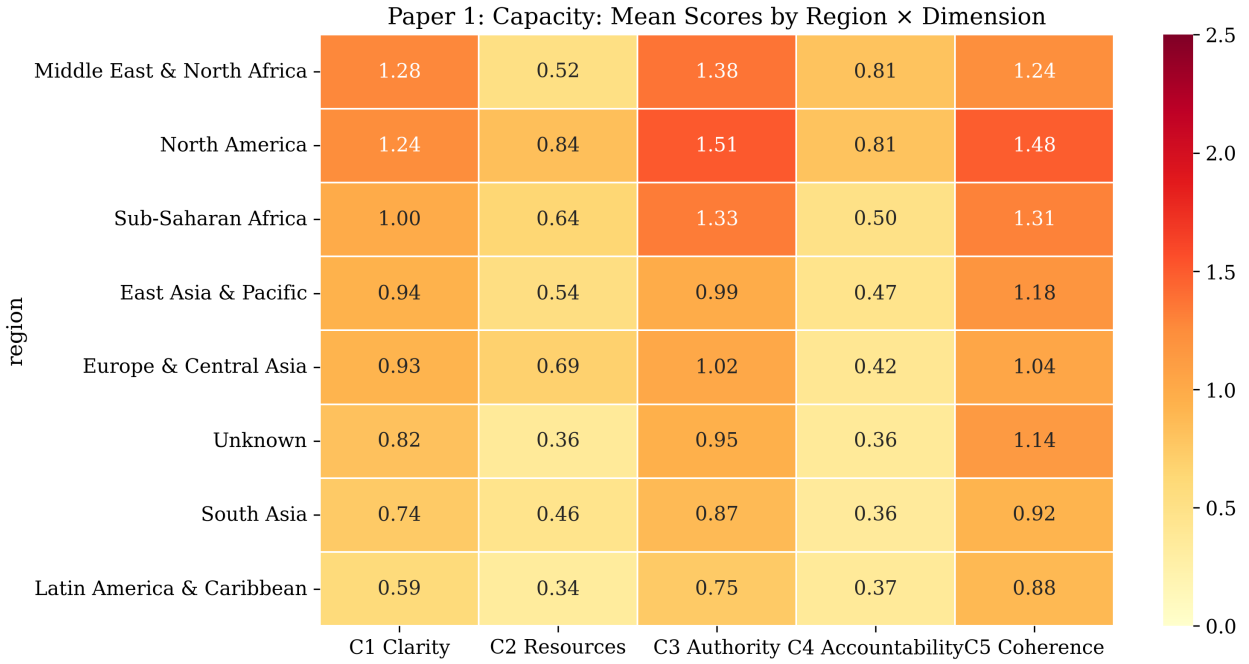

Figure 5.4: Heatmap of mean capacity scores by region and dimension. North America and Europe & Central Asia lead; Sub-Saharan Africa and South Asia trail.

Figure 5.4 shows substantial regional variation that does not map neatly onto income. **North America** leads across all dimensions, driven by the US and Canada. **Europe & Central Asia** has the broadest dimensional coverage, especially on Coherence (C5), reflecting the EU's multilevel coordination infrastructure.

More interesting are the developing regions that punch above their weight. **Latin America** scores above its income-group average on multiple dimensions, with particular strength in Authority (C3). Brazil, Colombia, and Argentina have adopted binding AI legislation with real enforcement teeth. This likely reflects the region's longstanding tradition of codified law (civil law systems), strong constitutional courts, and established data protection authorities—institutional infrastructure that predates AI governance and provides ready-made channels for algorithmic regulation. The Inter-American Development Bank and regional forums like the Red Iberoamericana have also facilitated policy coordination across the region.

Sub-Saharan Africa, while scoring lowest overall, shows surprising strength in Authority: Kenya, Rwanda, and South Africa have enacted AI-specific legislation with compliance mechanisms that exceed what many wealthy countries have managed. These countries benefit from institutional

investments in digital governance—Kenya's data protection framework, Rwanda's ICT-driven development strategy, South Africa's constitutional rights infrastructure—that provide foundations for AI-specific regulation. The African Union's AI strategy has also created a continental reference point, though adoption remains uneven.

East Asia & Pacific presents a mixed picture: high-performing jurisdictions like Singapore, Japan, and South Korea coexist with countries showing minimal AI governance activity. The region's diversity—spanning high-income technology leaders, rapidly industrialising economies, and small island states—makes regional averages particularly uninformative.

MENA and **South Asia** trail on most dimensions, reflecting later entry into AI governance (median first adoption in 2020–2021) rather than inherently weaker institutional capacity. The UAE and Saudi Arabia are notable exceptions in MENA, with comprehensive strategies driven by national economic diversification agendas.

Governance capacity reflects institutional and political factors well beyond simple wealth accumulation.

5.1.4 Policy-Type Variation

Not all policy documents carry the same implementation obligations. National strategies articulate visions; binding regulations establish legal requirements; ethics guidelines provide normative frameworks. These functional differences mean capacity scores should vary by type. This reflects the different purposes documents serve rather than genuine capacity differences across jurisdictions.

As Figure 5.5 confirms, **binding regulation** scores highest. Laws must specify enforcement mechanisms, responsible agencies, and monitoring procedures to be judicially enforceable. **National strategies** sit in the middle, strong on Clarity and Coherence but weaker on Resources and Authority (which strategies typically leave to implementing legislation). **Guidelines and principles** score lowest, reflecting their aspirational character. Ethics guidelines list desirable AI properties without specifying who implements them, how compliance is monitored, or what resources are allocated.

The pattern validates the measurement approach: scores reflect document content appropriate to policy type. It also means that jurisdictions relying on voluntary guidelines will systematically score lower than those with binding legislation, even if the voluntary approach works in practice. Subsequent analyses control for policy type accordingly.

5.1.5 Country Rankings and Correlation Structure

Which countries have the strongest AI governance capacity? Rankings compress the multidimensional framework into a single scale, losing nuance but gaining interpretive clarity. They reflect both the number of policies and their average quality. The top-scoring jurisdictions combine large portfolios with consistently high individual scores:

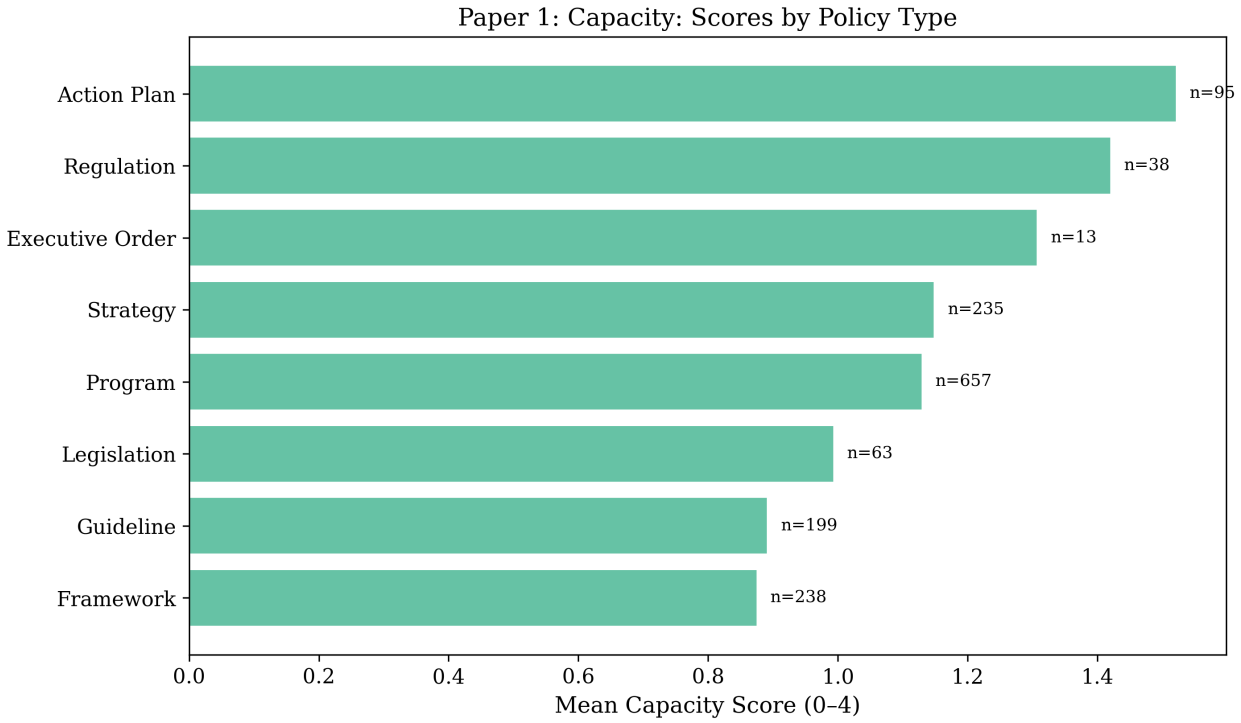

Figure 5.5: Capacity scores by policy type. Binding regulations score highest; guidelines and principles score lowest.

Table 5.4: Top 5 jurisdictions by capacity score

Rank	Jurisdiction	Mean Score	Income	N Policies
1	European Union	1.42	HI	60
2	Canada	1.38	HI	15
3	United Kingdom	1.32	HI	72
4	United States	1.28	HI	84
5	Colombia	1.21	UMI	8

The top four are unsurprising: the **EU** (1.42) leads on the strength of the AI Act and its complementary policies; **Canada** (1.38) and the **UK** (1.32) show that smaller portfolios can achieve high quality; the **US** (1.28) ranks fourth because its massive 84-document portfolio includes substantial variation across federal agencies and states.

The fifth slot is more interesting. **Colombia** (1.21), an upper-middle-income country, outperforms dozens of wealthy nations with just eight focused, well-designed policies. This is a preview of a pattern explored in Section 8.1: several developing countries, including Brazil (consistently top 10) and Kenya (above many European countries), achieve strong capacity through strategic design rather than resource abundance.

Correlation structure. The five capacity dimensions are positively correlated (r = 0.45 to 0.75)

but maintain sufficient discriminant validity to justify separate measurement. A common governance quality factor accounts for ~66% of variance in PCA, meaning policies that score high on one dimension tend to score high on others. However, the residual independence is substantively important: a policy can score high on Clarity (well-defined objectives) while scoring low on Resources (no budget), or achieve strong Coherence (coordination) despite weak Accountability (no monitoring). The strongest links—Authority with Coherence (r = 0.75) and Clarity with Authority (r = 0.70)—make structural sense: policies with legal mandates more readily establish coordination mechanisms.

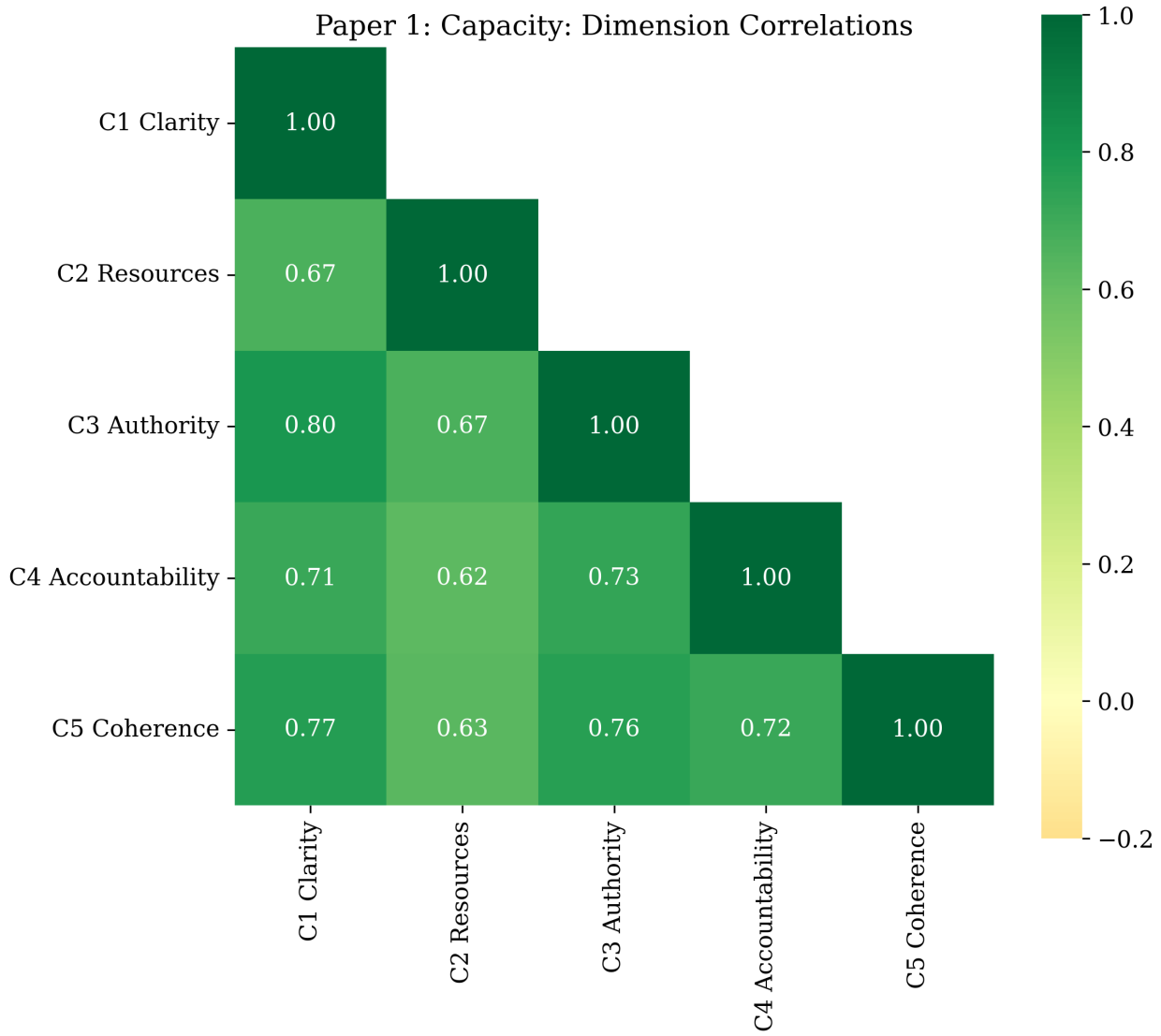

Figure 5.6: Correlation matrix across the five capacity dimensions.

6 Capacity Determinants

6.1 What Explains Governance Capacity?

The descriptive landscape revealed low capacity scores overall, modest income-group differences, and substantial within-group variation. This section moves from description to explanation, asking what drives variation in governance capacity across policies. Four complementary regression approaches—OLS, Tobit, quantile regression, and multilevel models—test whether national wealth, policy type, text quality, or institutional factors account for the observed patterns. The results consistently point to the same conclusion: GDP exerts a statistically significant but substantively negligible effect, while text quality and policy type emerge as the dominant predictors.

6.1.1 Baseline and Multilevel Models

The scatter in Figure 6.1 reveals the pattern prior to formal regression: the relationship with GDP is positive but weak, with many developing-country policies exceeding the fitted line and many high-income policies falling below it. The baseline OLS model:

$$\text{Capacity}_i = \beta_0 + \beta_1 \ln(\text{GDP}_{pc}) + \beta_2 \text{Year} + \beta_3 \text{Binding} + \beta_4 \text{GoodText} + \varepsilon_i$$

Table 6.1: OLS regression: capacity determinants ($R^2 = 0.436$, $N = 1{,}949$)

Variable	β	SE	t	p
Intercept	−0.536	0.245	−2.19	.029
log(GDP pc)	**0.086**	0.023	3.81	$< .001$
Year (centred)	0.010	0.006	1.75	.081
Binding regulation	**0.190**	0.069	2.73	.006
Good text quality	**1.004**	0.027	37.64	$< .001$

The results reveal three patterns. **Text quality dominates all substantive predictors**: the coefficient on `is_good_text` (= 1.004, t = 37.64) exceeds every policy-relevant variable by an order of magnitude. Policies with 500 words of extracted text score a full point higher on the 0–4 scale, primarily reflecting the greater evidentiary basis available to LLM coders. This represents largely a measurement artifact rather than a governance finding.

GDP demonstrates statistical significance but substantive weakness (= 0.086). Tripling a country's wealth from $10K to $30K per capita raises predicted capacity by less than one-tenth

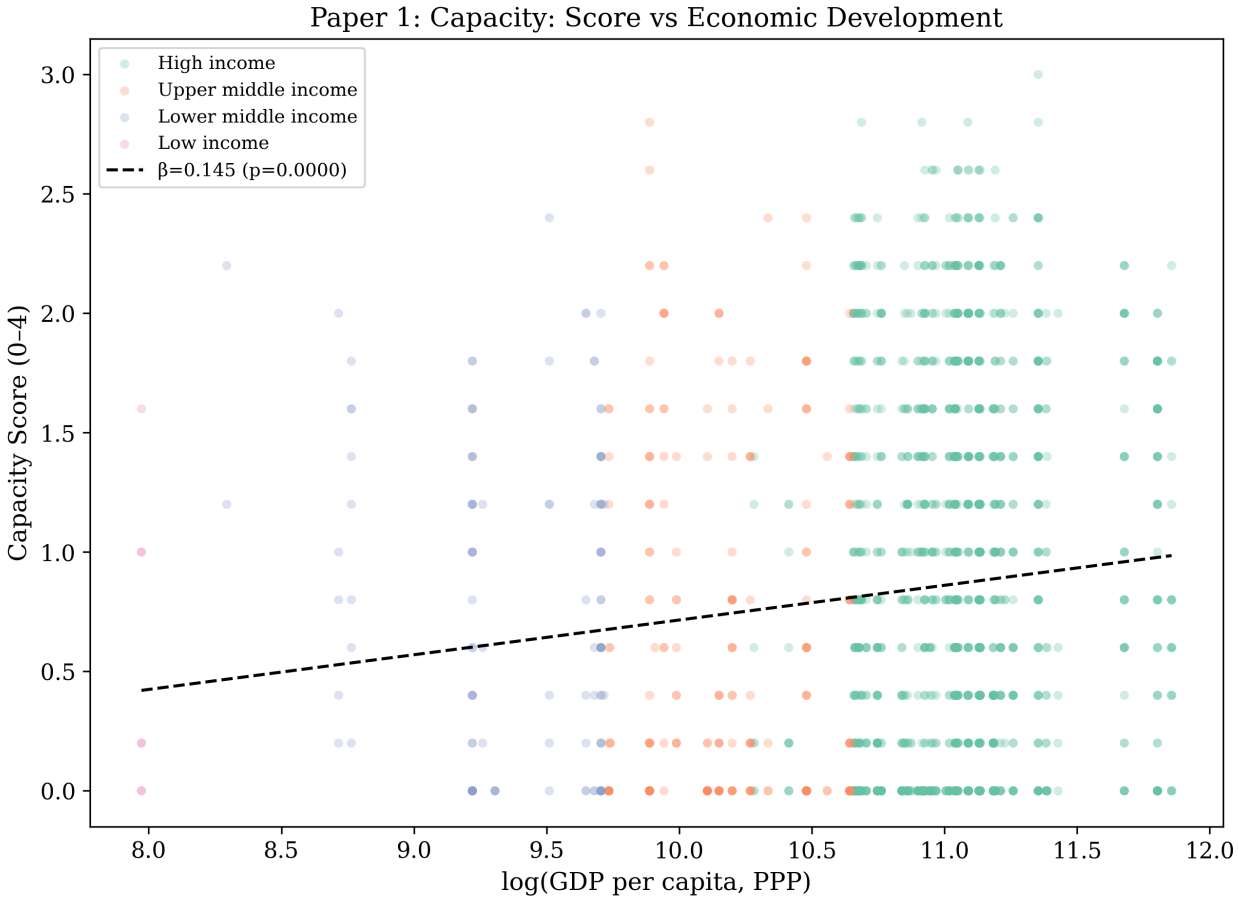

Figure 6.1: Scatter plot of capacity scores against log GDP per capita, with regression line. The relationship is positive but weak.

of a scale point. Finally, **binding regulation adds approximately 0.19 points** (= 0.190), reflecting the structural requirement that laws specify authorities and enforcement mechanisms.

The model's R^2 of 0.436 appears respectable until text quality is removed—the explained variance then collapses to 0.012. GDP, year, and document type *combined* explain only 1.2% of capacity variance. Nearly all explanatory power derives from the measurement artifact of text length.

Multilevel models. OLS assumes independence across observations, but policies nest within countries. Multiple policies from a single jurisdiction share institutional features, creating dependency that inflates standard errors. A random-intercept multilevel model addresses this structure:

$$\text{Capacity}_{ij} = \gamma_0 + \gamma_1 \ln(\text{GDP}_{pc,j}) + u_j + \varepsilon_{ij}$$

where $u_j \sim N(0, \sigma_u^2)$ is the country random effect.

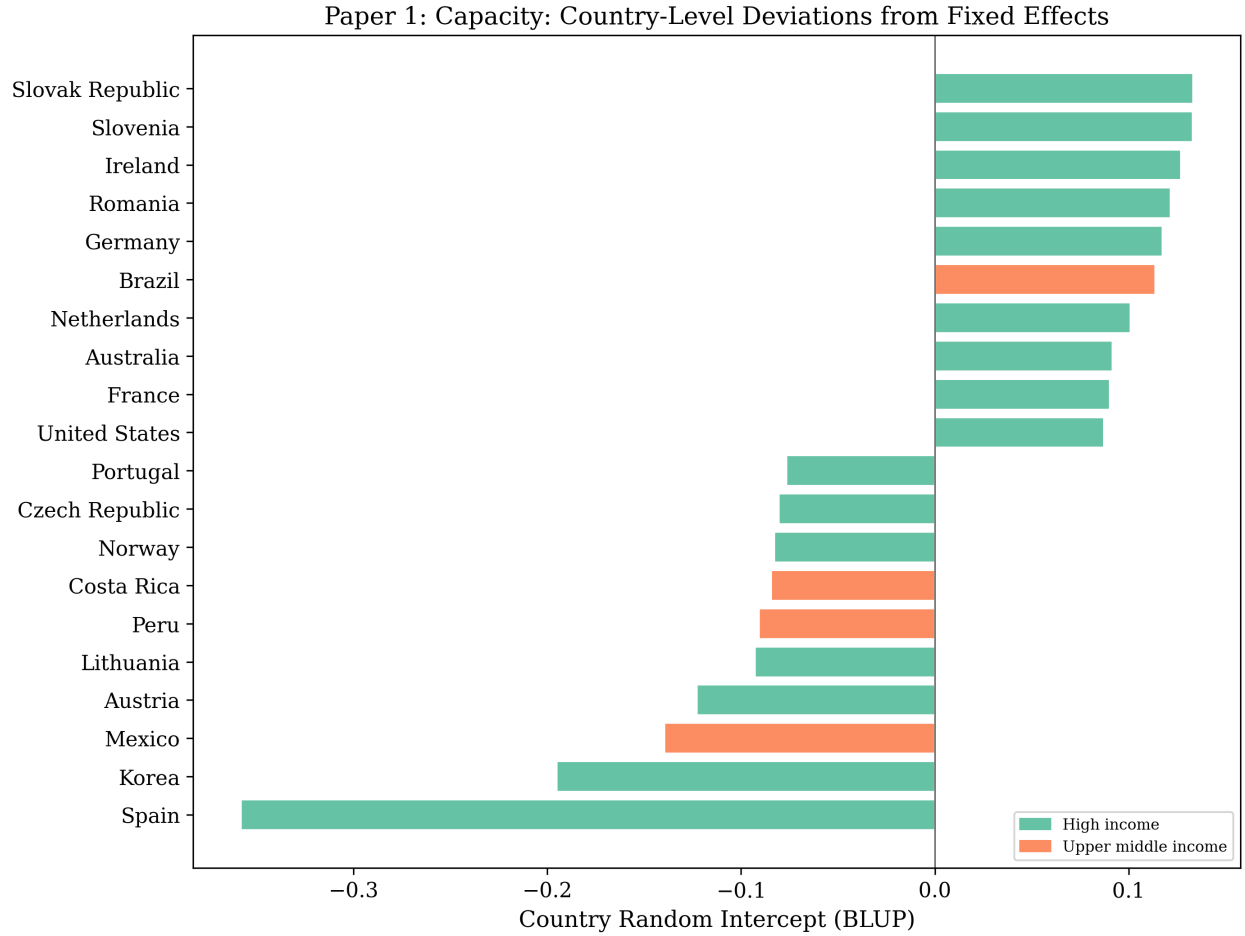

Figure 6.2: Country random effects from the multilevel capacity model. Most countries cluster near zero, with a few clear over- and under-performers.

Table 6.2: OLS vs. multilevel model comparison

Metric	OLS	Mixed
GDP β	0.088	**0.066**
GDP p	$< .001$	.038
Country ICC	—	**0.091**
LR test vs OLS	—	$p = .007$

The **country-level ICC of 0.091** reveals the key finding: only 9.1% of total variance occurs between countries. The remaining 90.9% represents variation between policies *within* countries. Policies from a single country differ more from each other than the average policy from that country differs from policies elsewhere.

Accounting for clustering reduces the GDP coefficient by approximately 25% (0.088 → 0.066), as OLS double-counts country-level variation. The likelihood ratio test confirms the multilevel spec-

ification is statistically preferred (p = .007), though the practical conclusion remains unchanged: GDP effects prove even smaller than OLS suggested.

6.1.2 Distributional Analysis

OLS and multilevel models estimate *mean* effects. However, GDP might prove consequential for establishing basic infrastructure (lower quantiles) while remaining irrelevant for distinguishing strong from excellent policies (upper quantiles)—or the reverse. Quantile regression (Koenker and Bassett 1978) permits testing across the distribution.

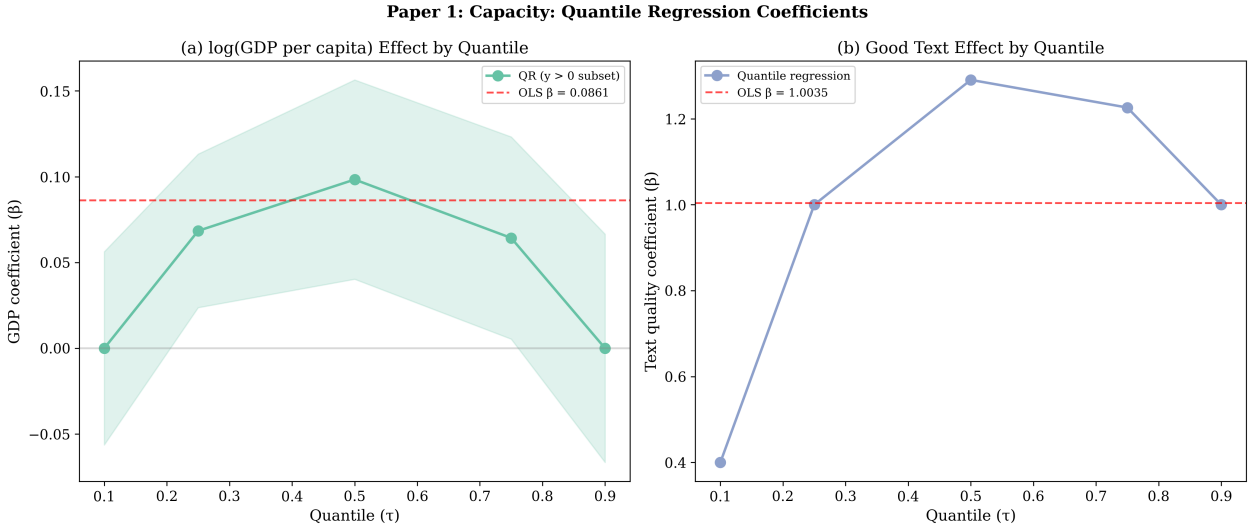

Figure 6.3: Quantile regression coefficients for GDP across the capacity distribution. GDP matters at the median but not at the extremes — an inverted-U pattern.

Figure 6.3 visualizes the GDP effect across quantiles, revealing a clear **inverted-U pattern** where GDP matters most at the median and least at both tails. The formal estimates quantify this heterogeneity:

Table 6.3: Quantile regression: GDP effect across the capacity distribution

Quantile (τ)	GDP β	SE	p
0.25 (positive subset)	0.068	0.024	.005
0.50	**0.098**	0.019	$< .001$
0.75	0.064	0.029	.028
OLS (reference)	0.086	0.023	$< .001$

Figure 6.3 shows an **inverted-U**: GDP matters most at the median (= 0.098) and less at both tails (= 0.064–0.068). At the bottom, many policies score zero regardless of GDP—no amount of national wealth transforms a press release into an implementation plan. At the top, excellent policies emerge from factors orthogonal to wealth: policy learning, political commitment, design

sophistication. Brazil's comprehensive framework did not require first-world GDP. Wealth helps countries avoid governance failures (moving from zero to moderate) but does not ensure excellence.

Tobit regression. The 27.6% floor effect (scores censored at zero) biases OLS downward—scores cannot go below zero, so OLS treats observed zeros as the "true" capacity when they might reflect even lower latent quality. Tobit (Tobin 1958) models this left-censoring explicitly:

$$\text{Capacity}_i^* = \mathbf{x}_i'\beta + \varepsilon_i, \quad \text{Capacity}_i = \max(0, \text{Capacity}_i^*)$$

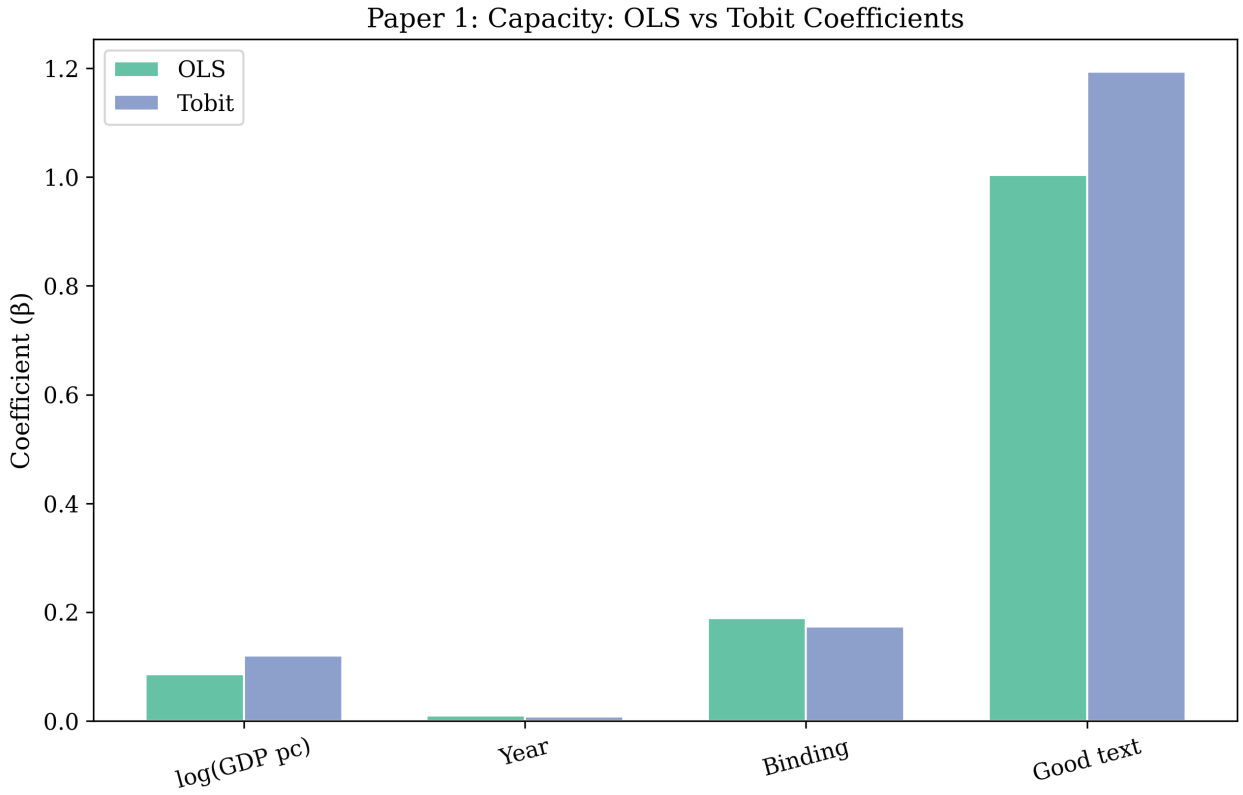

Figure 6.4: Comparison of OLS and Tobit coefficients for the capacity model. Tobit coefficients are systematically larger, reflecting correction for censoring.

Table 6.4: OLS vs. Tobit comparison

Variable	OLS β	Tobit β	Ratio
log(GDP pc)	0.086	**0.121**	1.41×
Year	0.010	0.008	0.80×
Binding regulation	0.190	**0.174**	0.92×
Good text quality	1.004	**1.193**	1.19×
σ	—	0.742	—

The Tobit GDP coefficient (= 0.121) is **41% larger** than OLS (0.086), confirming that OLS understates the true wealth-capacity relationship when floor effects are present. But even corrected, 0.121 remains modest: tripling GDP per capita adds 0.12 points on a 4-point scale. The correction matters methodologically without changing the substantive conclusion.

6.1.3 Synthesis

Across four estimation strategies, three findings hold.

Text quality dominates. In every specification, document length dwarfs GDP, policy type, and time trends. Much of what looks like governance quality is actually documentation quality. The robustness checks in Section 9.1 test whether anything survives once measurement quality is held constant.

GDP matters, but barely. After multilevel correction for clustering (−25%) and Tobit correction for censoring (+41%), GDP per capita explains 3–12% of capacity variance depending on specification. Tripling national wealth adds 0.07–0.12 points on a 4-point scale. Institutional choices and policy design matter more.

Binding regulation helps. Laws consistently outscore voluntary guidelines by ~0.19 points, reflecting structural necessities of legal enforceability. Countries can strengthen implementation readiness by choosing binding instruments—a strategic option available regardless of wealth.

The multilevel decomposition makes the point clearly: 91% of variance is *within* countries. Seeking country-level explanations—GDP, democracy, institutional quality—inevitably leaves most variation unexplained. What matters is the individual policy: its type, how well it is documented, and what design choices it makes.

7 Capacity Inequality and Clusters

7.1 Within vs. Between: Decomposing the Governance Gap

The regression analysis established that GDP explains remarkably little variance in governance capacity. This section examines the structural distribution of that variation: does the governance gap primarily separate income groups, or does most inequality reside within them? Using Gini coefficients, Lorenz curves, and Theil decomposition, the analysis reveals that nearly all variation in capacity scores occurs within income groups rather than between them—a finding that fundamentally reframes the conventional narrative of a global governance divide.

7.1.1 Inequality Decomposition

Two scenarios are consistent with weak GDP effects: (1) a small but stable income-group gap with most variation within countries, or (2) a substantial gap obscured by enormous within-group heterogeneity. Distinguishing these requires formal inequality decomposition.

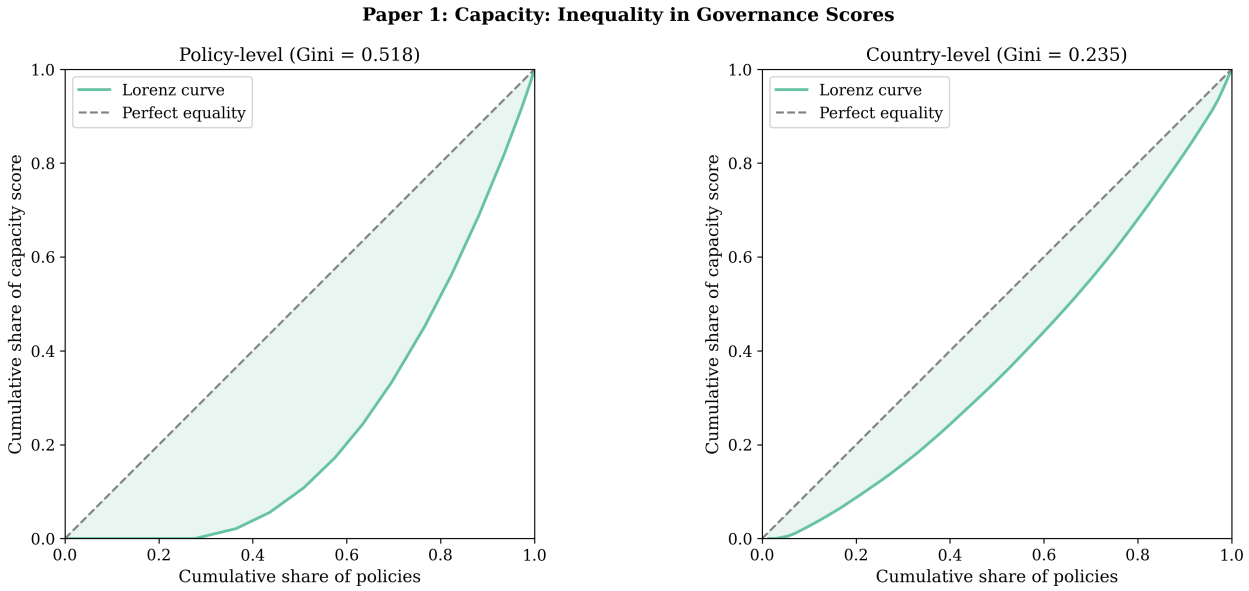

Figure 7.1: Lorenz curves for capacity scores, by income group and overall. Both groups exhibit substantial inequality, with developing countries showing slightly more concentration.

Figure 7.1 presents Lorenz curves visualizing inequality within and between income groups. If capacity were equally distributed, policies would fall along the 45-degree line where each percentile

of policies accounts for its proportional share of total capacity. The observed curves bow below this line, indicating concentration: top-scoring policies capture disproportionate shares of total capacity. The Gini coefficient quantifies this departure from perfect equality:

Table 7.1: Gini coefficients for capacity scores

Metric	Value
Gini (all countries)	0.518
Gini (HI only)	0.499
Gini (Developing)	0.593
Gini (country means)	0.235

The overall Gini of **0.518** indicates substantial inequality, comparable to income inequality in moderately unequal societies. The pattern of concentration proves noteworthy: **developing countries exhibit higher within-group inequality (0.593) than high-income countries (0.499)**. The range among developing countries, from high-capacity implementations in Brazil and Colombia to minimal AI governance frameworks elsewhere, exceeds the range observed among wealthy nations.

If developing countries demonstrated uniformly weak capacity, their within-group Gini would be low. Instead, the coefficient of 0.593 indicates substantial heterogeneity. The aggregate label "developing country" therefore obscures more variation than it reveals, conflating high-performing jurisdictions with countries exhibiting minimal AI governance infrastructure.

The country-means Gini (0.235) is less than half the overall (0.518), confirming that aggregating to the country level eliminates much variation. Most inequality lives at the policy level within countries.

Theil decomposition. While the Gini coefficients suggest within-group inequality dominates, precise quantification requires decomposition. Theil's T index permits exact partitioning into between-group and within-group components.

Figure 7.2 visualizes the decomposition. The formal results prove more pronounced than Gini coefficients suggested:

Table 7.2: Theil decomposition of capacity inequality

Component	Share
Between income groups	**1.2%**
Within income groups	**98.8%**

The decomposition reveals that **income-group classification explains 1.2% of total capacity inequality**. The remaining **98.8% resides within income groups**. Knowledge of whether a country is classified as "high-income" or "developing" provides minimal information about its governance capacity.

To contextualize this finding: variation between Luxembourg's sophisticated framework and less developed European governance systems exceeds the average gap between income groups. Afghanistan

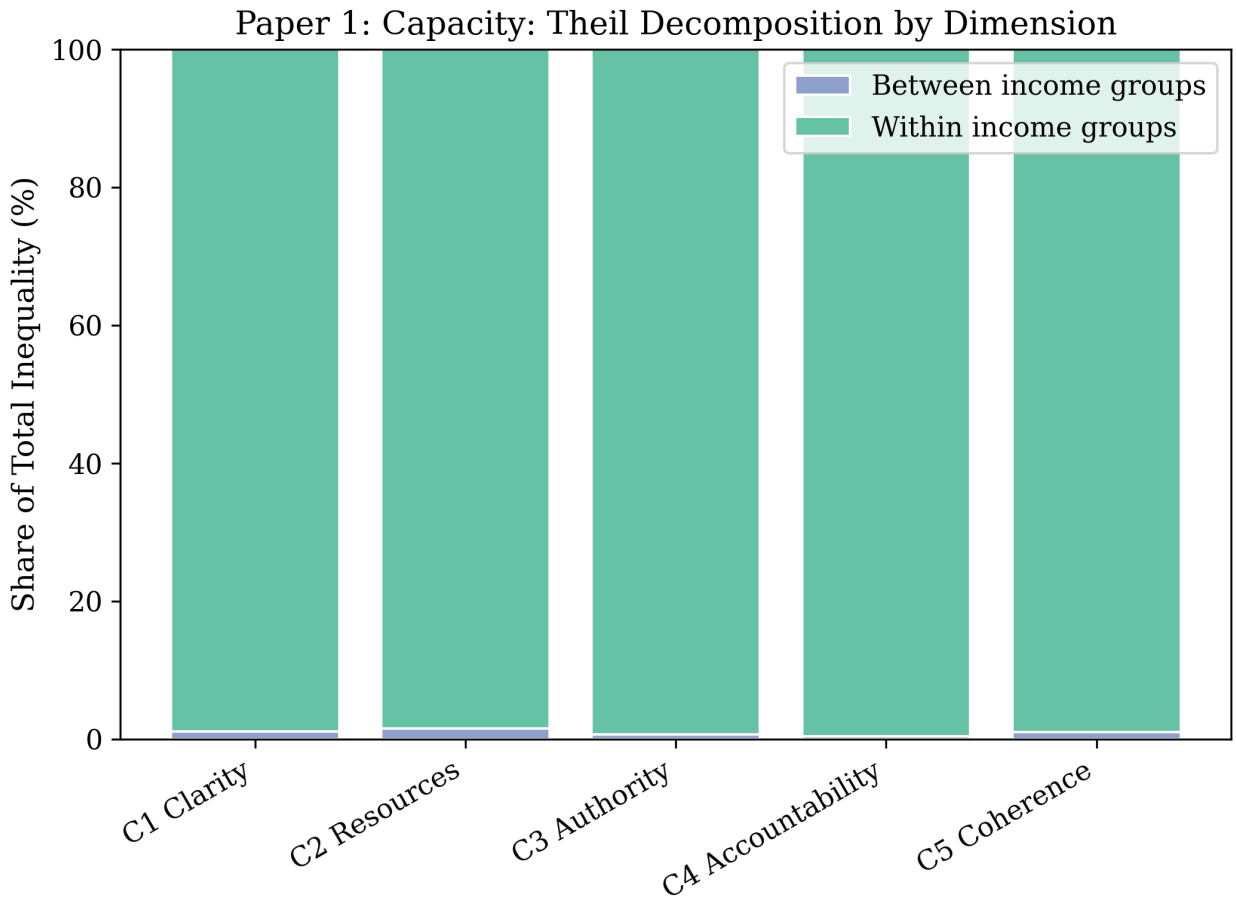

Figure 7.2: Theil decomposition of capacity inequality: 98.8% within groups, only 1.2% between income groups.

and Brazil both carry the "developing" classification, yet their governance capacity differs by more than the average cross-group difference.

The same pattern appeared in the multilevel analysis (91% within-country variance; Section 6.1), but the concentration here is even more extreme. The policy implication is direct: development interventions targeting "developing countries" as a category will misallocate resources, assisting Brazil and Colombia (which have already achieved strong capacity) while potentially neglecting wealthy countries with weak frameworks.

7.1.2 Policy Portfolios and Governance Typologies

Is the capacity gap about *breadth* (developing countries missing entire dimensions) or *depth* (covering the same dimensions but with less detail)? The distinction matters: if countries miss dimensions, interventions should ensure comprehensive frameworks; if they cover all dimensions shallowly, the focus should be on strengthening existing areas.

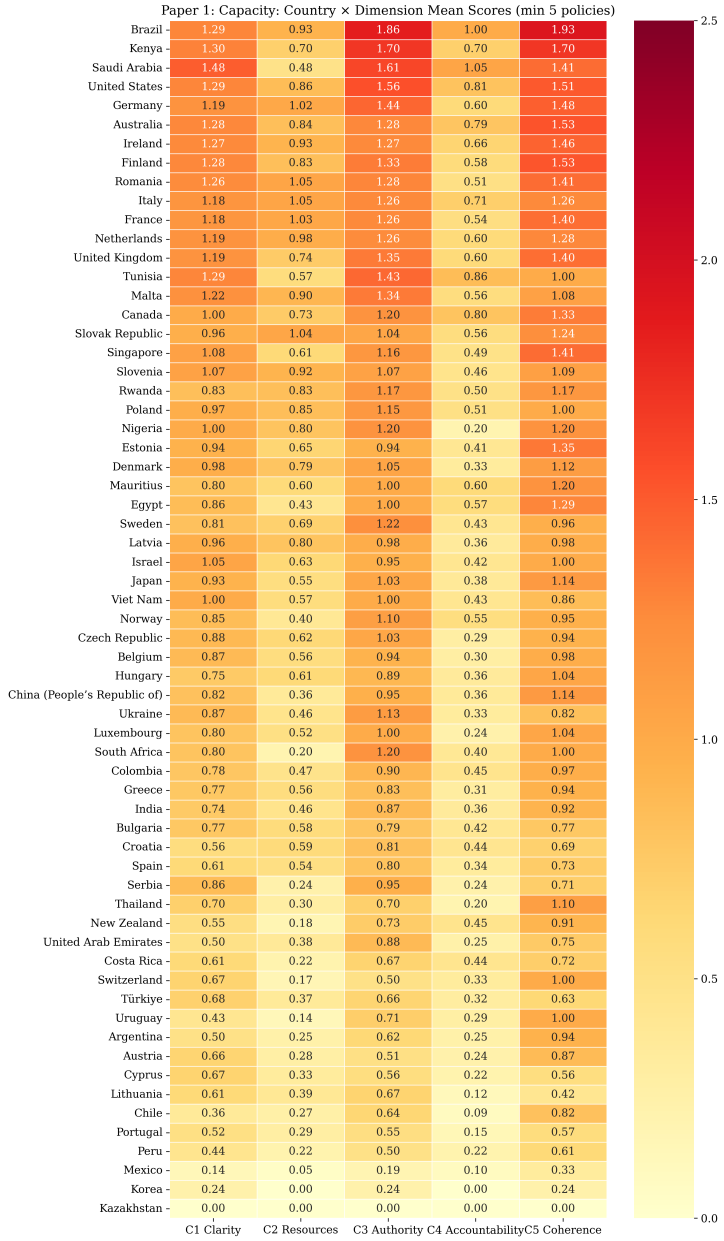

Figure 7.3: Heatmap of policy portfolio coverage by country and dimension. Most countries cover all five dimensions in at least one policy.

Figure 7.3 maps policy portfolio coverage across countries and dimensions, revealing that most jurisdictions — regardless of income — address all five capacity dimensions in at least some policy. The heatmap shows dense coverage rather than systematic gaps, suggesting that breadth deficits prove less consequential than depth variations.

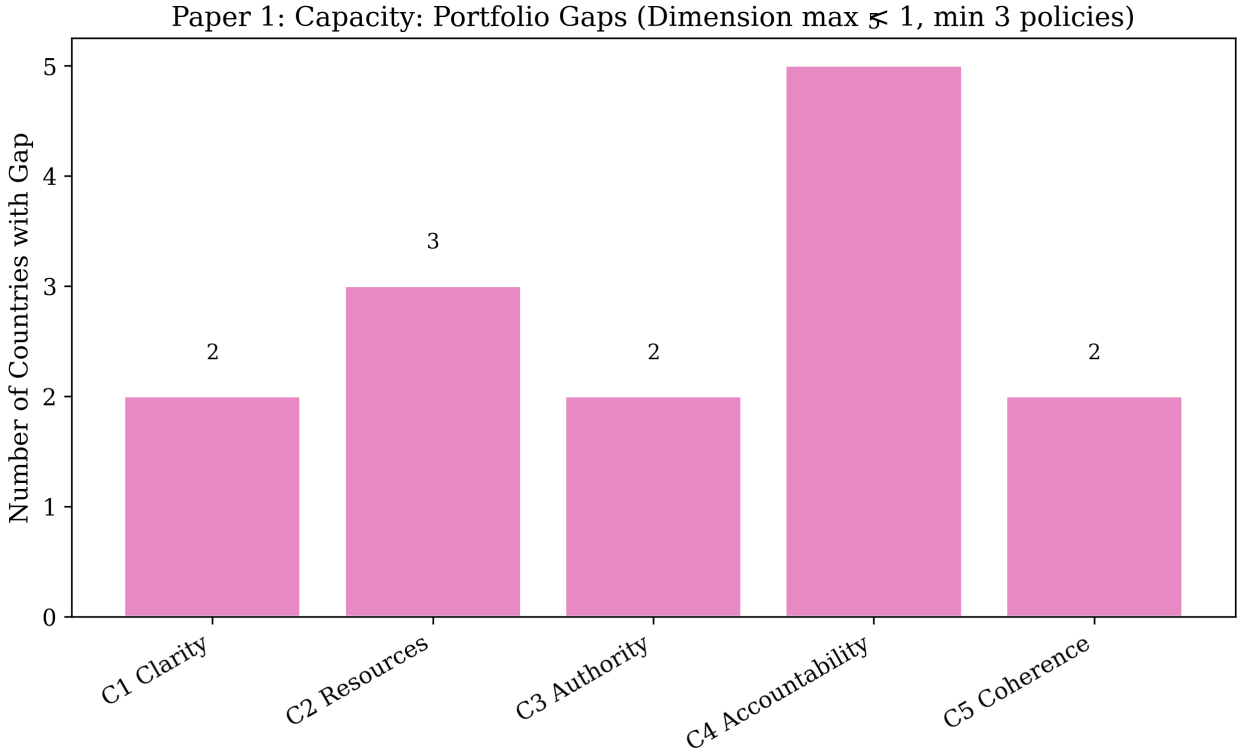

Figure 7.4: Portfolio breadth gap between income groups, by dimension. C4 Accountability shows the largest coverage gap.

Figure 7.4 quantifies dimensional coverage differences between income groups, revealing that gaps concentrate in specific dimensions rather than reflecting uniform undercoverage:

Table 7.3: Policy portfolio breadth by income group

Metric	HI	Developing	p
Mean breadth (out of 5)	4.95	4.52	.137
Countries with 5/5 coverage	93%	—	—
Least covered dimension	C4 (92.6%)	C4 (92.6%)	—

The breadth gap is **not statistically significant** (p = .137). High-income countries average 4.95 dimensions out of 5; developing countries average 4.52. Both groups cover most dimensions. And both groups share the same weakest link: **C4 Accountability** is the most commonly missing dimension across both groups (92.6% coverage).

So the capacity gap is **about depth, not breadth**. Developing countries recognize the same governance requirements; they just provide less implementation detail within each dimension.

K-means clustering. Do distinct governance *typologies* exist, recognizable archetypes with characteristic dimensional profiles? I use K-means clustering to find out, with optimal k determined by silhouette analysis.

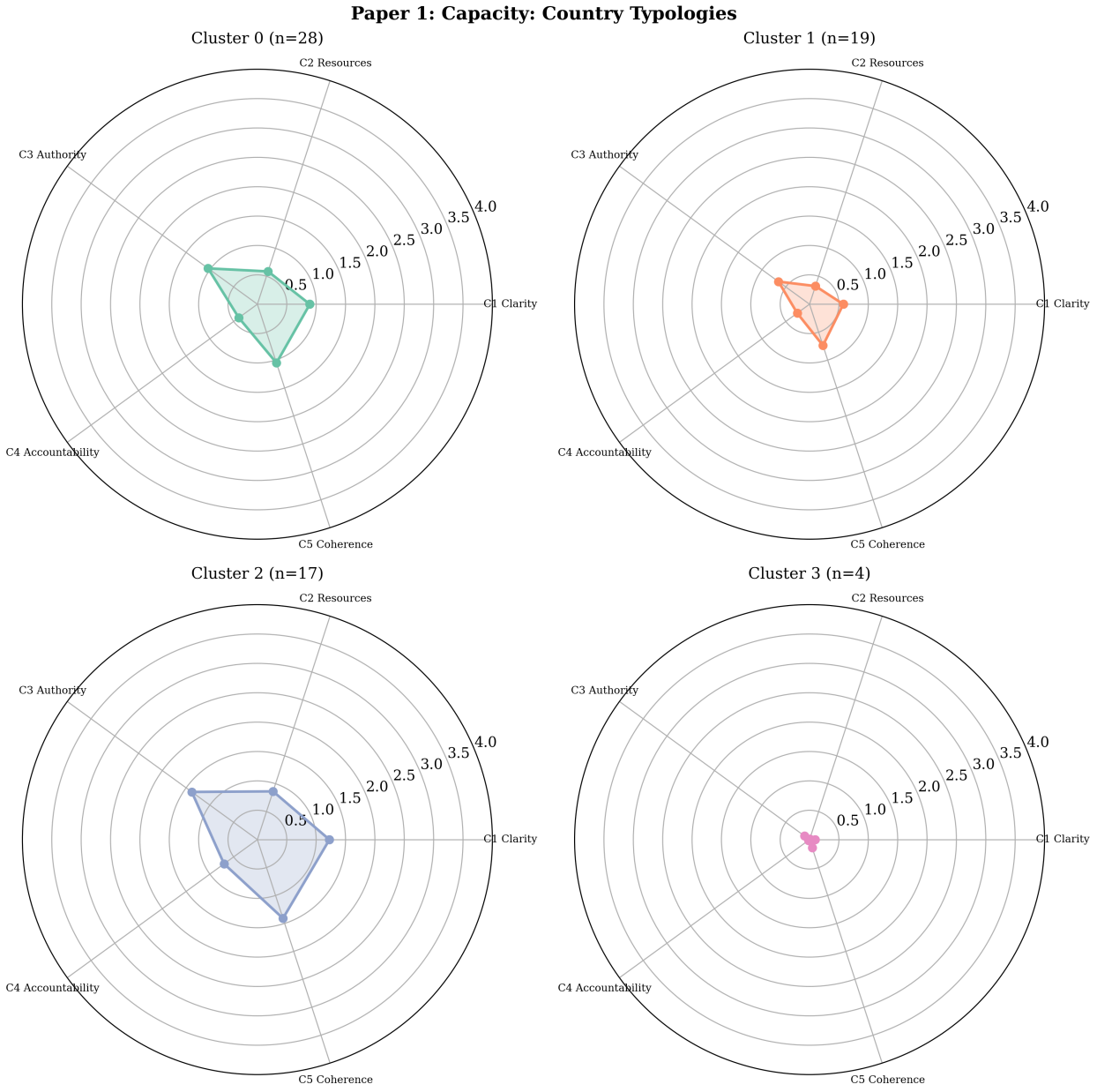

Figure 7.5: Radar chart of capacity dimension profiles for the two identified clusters. Cluster 1 ("Low Capacity") shows uniformly low scores; Cluster 2 ("Moderate Capacity") shows elevation across all dimensions.

Two clusters emerge (k = 2, silhouette = 0.41). They differ in *level* rather than *shape*: both show the same relative strengths and weaknesses, just at different absolute heights.

Cluster 1 ("Low Capacity", ~60% of policies): Uniformly low scores (0–1 range). Aspirational statements, preliminary frameworks, brief announcements. Both income groups contribute.

Cluster 2 ("Moderate Capacity", ~40%): Elevated across dimensions (1.5–3 range), with the usual weaknesses in Resources and Accountability. Again, both income groups are represented.

The fact that **both clusters contain policies from both income groups** reinforces the main point: the distinction between low-capacity and moderate-capacity governance matters more than the distinction between rich and poor countries. The k = 2 solution (stable under bootstrap; see Section 9.1) suggests AI governance capacity operates as a continuum from absent to present, not as distinct regulatory philosophies.

8 Capacity Dynamics

8.1 Temporal Trends, Diffusion, and the Efficiency Frontier

The preceding sections established the cross-sectional landscape: low average capacity, weak GDP effects, and inequality concentrated within income groups. This section introduces a temporal dimension, asking whether these patterns are stable, converging, or diverging over the 2017–2025 period. It then examines policy diffusion mechanisms—how governance approaches spread across jurisdictions—and identifies countries that achieve governance capacity exceeding expectations given their economic development level.

8.1.1 Temporal Trends

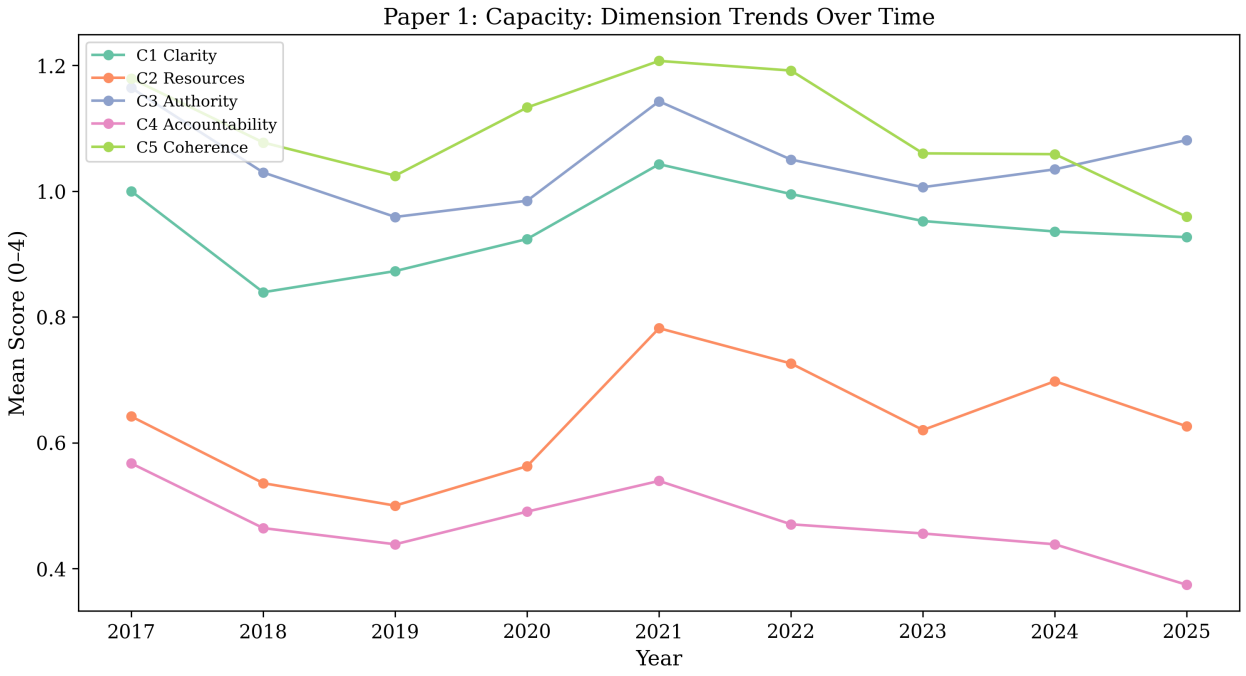

Figure 8.1: Capacity dimension scores over time (2017–2025). Most dimensions show modest upward trends.

Individual dimensions trend modestly upward from 2017 to 2025, but the income-group gap stays flat: parallel lines, no convergence, no divergence:

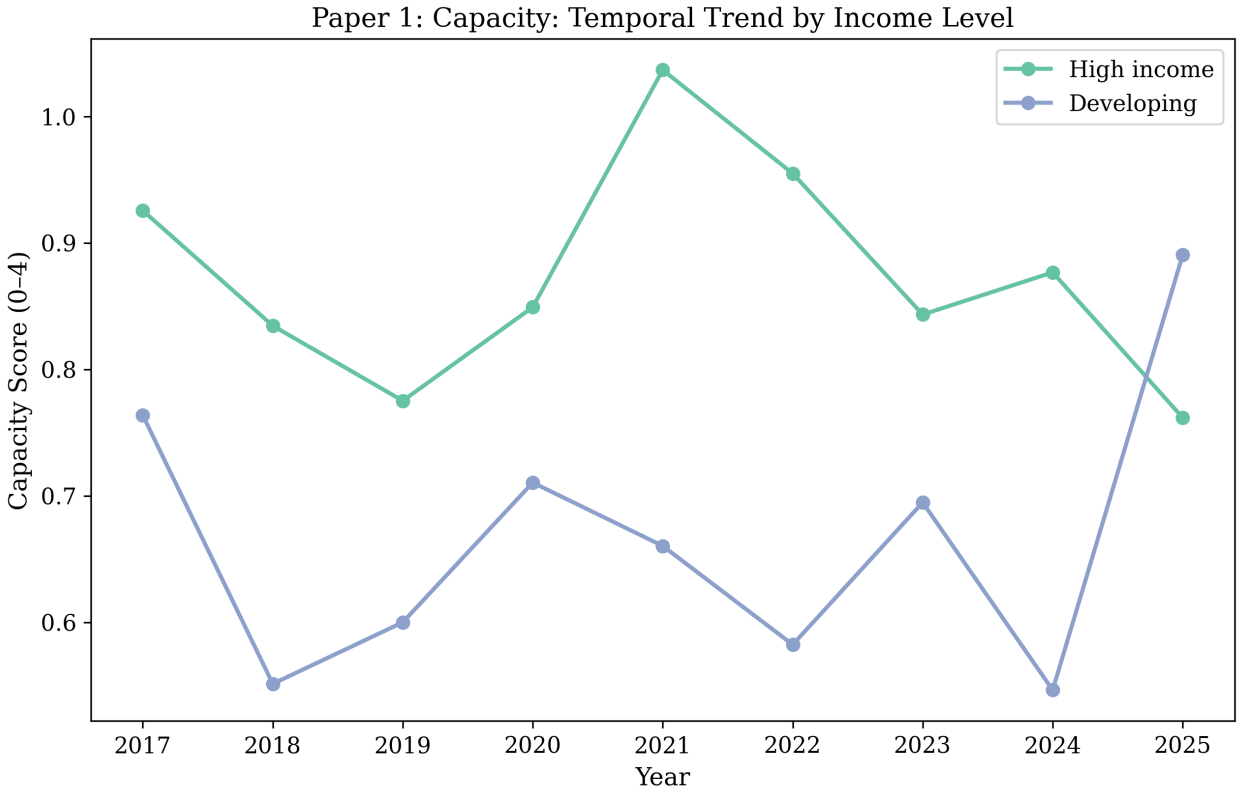

Figure 8.2: Capacity trends by income group over time. The gap between HI and developing countries remains stable.

Table 8.1: Convergence test for capacity scores

Metric	Capacity
Income × Year interaction	$\beta = +0.0003$, $p = .98$
HI temporal slope	−0.0001/yr (n.s.)
Developing temporal slope	+0.010/yr (n.s.)
Gap trend	Stable

No convergence is detected. The interaction coefficient (= +0.0003, p = .98) proves statistically indistinguishable from zero. High-income countries remain flat (−0.0001/yr, n.s.); developing countries show modest improvement at +0.010/yr (n.s.), insufficient to narrow existing gaps.

This pattern contrasts with ethics scores (see the ethics dynamics analysis), where significant convergence occurs through high-income countries *declining*. Capacity resists this pattern. The likely explanation: ethical principles can be adopted through citation of international frameworks, whereas building implementation infrastructure (budgets, staff, enforcement mechanisms) requires institutional development that policy learning alone cannot substitute. An alternative interpretation: if text quality improves over time (newer policies better documented), apparent stability could mask underlying changes. The robustness checks in Section 9.1 address this possibility.

Figure 8.3 and Figure 8.4 present the detailed convergence analysis across dimensions and income groups, confirming the pattern of temporal stability.

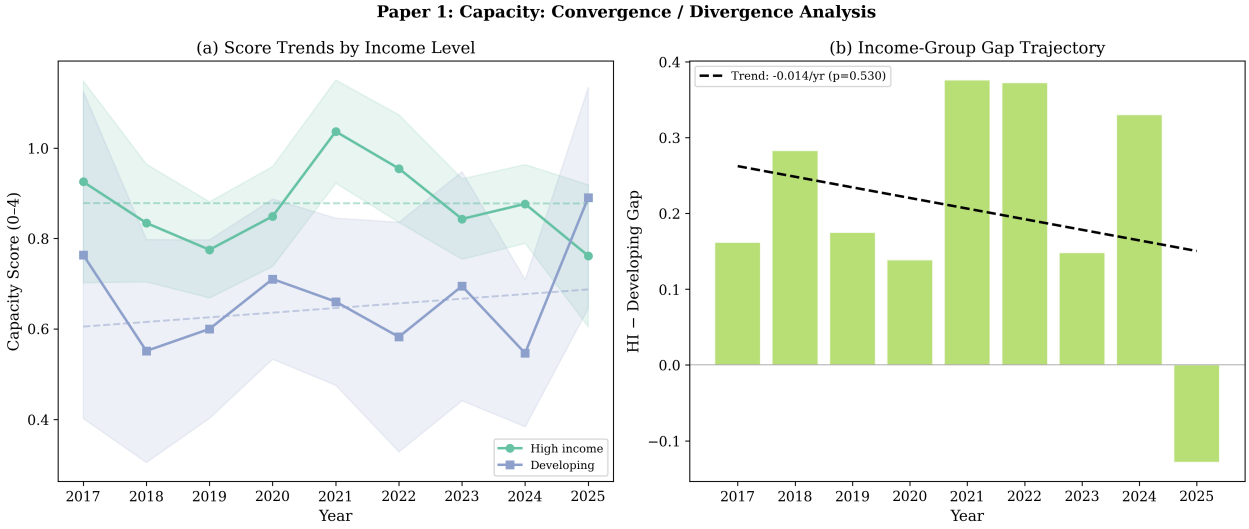

Figure 8.3: Detailed convergence analysis showing income-group trends and gap evolution for capacity dimensions.

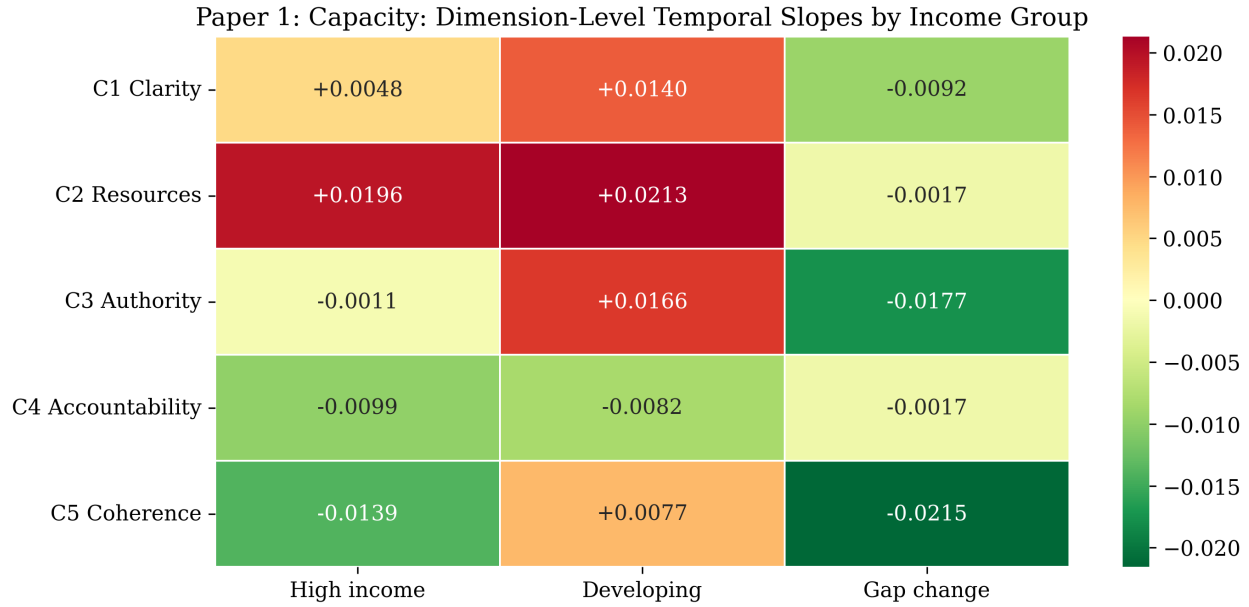

Figure 8.4: Dimension-level convergence patterns. No dimension shows statistically significant convergence.

8.1.2 Diffusion and the Efficiency Frontier

The mechanisms of policy spread warrant examination. The comparative policy literature proposes two models. **Vertical diffusion** (Bradford's (2020) "Brussels Effect"): innovations cascade from wealthy jurisdictions to developing countries through aid, technical assistance, and emulation. **Horizontal diffusion**: countries learn from peers facing similar constraints. The distinction carries policy implications: vertical diffusion suggests that strengthening high-income frameworks will eventually benefit developing countries; horizontal diffusion implies South-South cooperation proves more consequential.

This analysis tracks when each country first adopted AI governance policies and classifies diffusion direction.

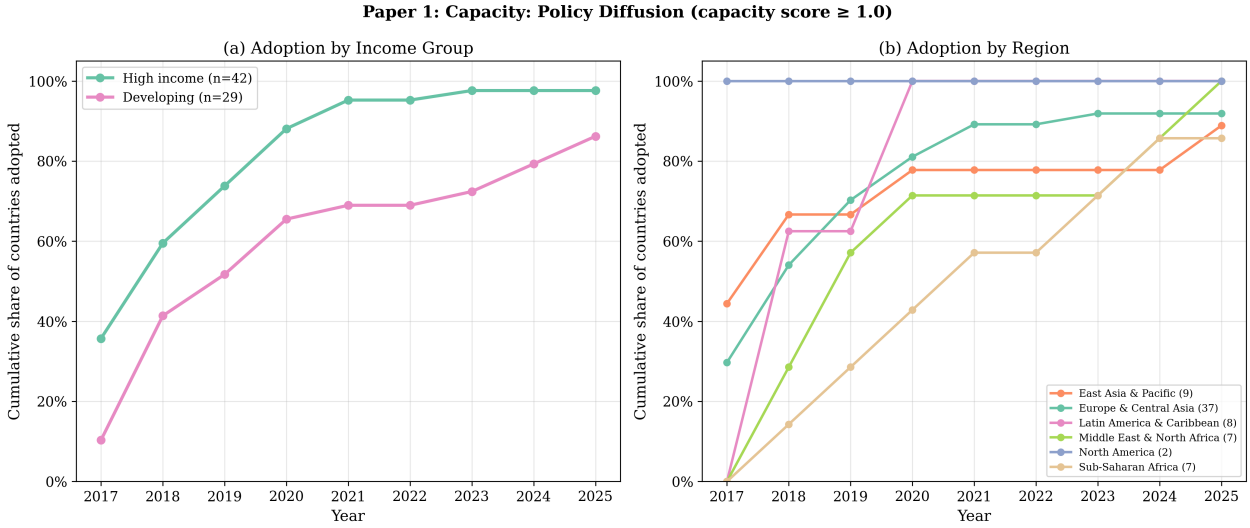

Figure 8.5: Cumulative adoption curves by income group and region. HI countries adopted ~1.3 years earlier, but diffusion is overwhelmingly horizontal.

The adoption curves show high-income countries reaching critical mass slightly earlier, but both groups surge in 2018–2020 in parallel, not sequential patterns:

Table 8.2: Policy diffusion patterns for capacity

Metric	Value
HI median first adoption	2018
Developing median first adoption	2019
Adoption lag (HI earlier by)	**1.3 years** ($p = .030$)
HI adoption by 2025	98%
Developing adoption by 2025	86%
Diffusion direction	**98% horizontal**

First, **high-income countries adopt about 1.3 years earlier** (median 2018 vs. 2019, p = .030), statistically significant but substantively modest. By 2025, 98% of high-income and 86% of

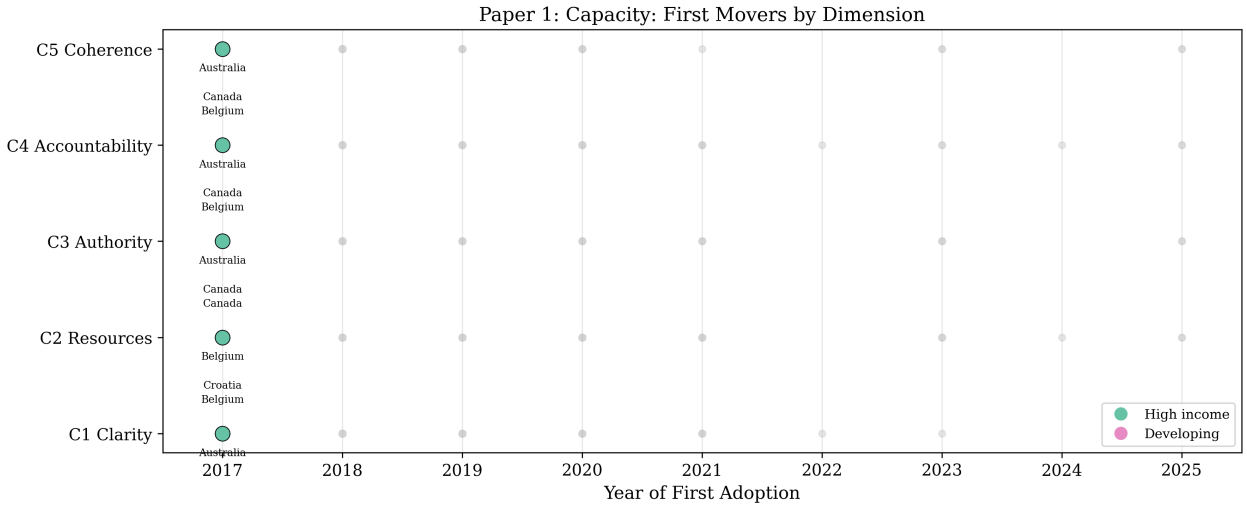

Figure 8.6: First movers in AI governance capacity, plotted by adoption year and income group.

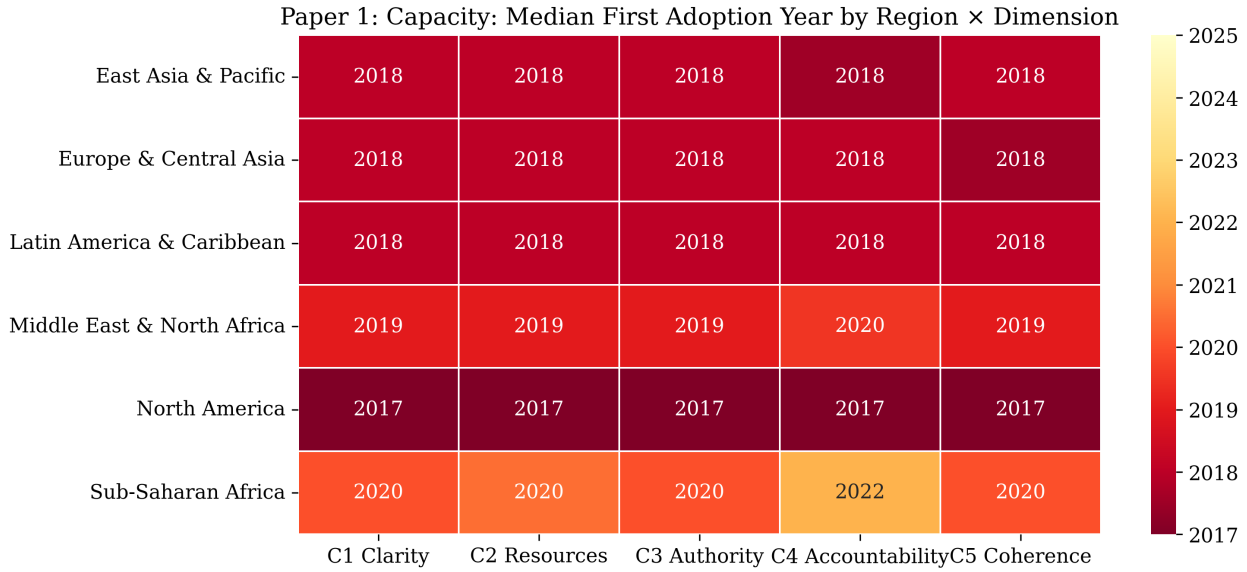

Figure 8.7: Regional diffusion heatmap showing the spread of AI governance policies across regions and years.

developing countries have adopted at least one AI governance policy. The gap is about timing, not whether countries adopt at all.

Second, and more important: **98% of diffusion is horizontal.** Countries learn from peers at similar income levels, not from wealthy-country frameworks cascading downward. This challenges the Brussels Effect hypothesis as applied to AI governance. I find minimal evidence of developing countries copying GDPR-inspired data protection or EU AI Act risk-based approaches. Instead, they build policies around their own institutional priorities and regulatory traditions.

Why? Wealthy-country models assume infrastructure that developing countries lack: data protection authorities, judicial capacity for algorithmic accountability, technical audit expertise. Direct transplantation is infeasible. Countries look to peers who adapted governance to similar resource constraints, or they resist perceived regulatory imperialism by developing indigenous approaches.

The mechanisms of horizontal diffusion are visible in the data. **Regional bodies** serve as transmission channels: the African Union's Continental AI Strategy provided a shared template that Kenya, Rwanda, and South Africa adapted to national contexts; the Red Iberoamericana de Inteligencia Artificial connected Latin American policymakers drafting national strategies simultaneously; and ASEAN's governance framework circulated among Southeast Asian countries with varying levels of AI maturity. **Bilateral technical assistance** between peers—Brazil advising Colombia, Singapore supporting ASEAN neighbours—transfers governance knowledge without the power asymmetry of North-South aid. **International conferences and working groups** (the Global Partnership on AI, the AI for Good summits, OECD committee meetings) create spaces where officials from similar-income countries compare approaches, often discovering that peers have solved problems they are facing. **Policy entrepreneur networks**—individuals who have drafted governance frameworks in one country and then advise neighbouring jurisdictions—facilitate direct knowledge transfer within income tiers.

The regions with the largest adoption lag, **Sub-Saharan Africa and MENA** (14–29% adoption by 2019 versus 100% in North America), lack these dense policy networks. They have fewer regional coordination mechanisms, fewer shared-language policy communities, and fewer international conference participants. Yet even these late adopters diffuse horizontally when they do adopt, suggesting that the constraint is network density rather than any inherent barrier to governance innovation.

Governance efficiency frontier. GDP explains little, and diffusion is horizontal. But which specific countries punch above their weight? I use Free Disposable Hull (FDH) analysis to construct an efficiency frontier, connecting countries that achieve higher governance scores than any other country at similar or lower GDP.

Countries above the regression line in Figure 8.8 outperform GDP predictions; those below underperform. Figure 8.9 ranks them by residual distance:

Table 8.3: Efficiency frontier results for capacity

Metric	Value
OLS R^2 (score ~ GDP)	**0.035**
Top overperformer	Brazil (+0.69)
Top underperformer	Kazakhstan (−0.75)
Frontier countries (FDH)	Uganda → Rwanda → Kenya → Brazil
Most efficient (score/$10k GDP)	Rwanda (3.10), Kenya (1.91)
Mean distance to frontier	0.588

GDP explains 3.5% of country-level capacity variation ($R^2 = 0.035$). The scatterplot disperses widely around the regression line. Knowing a country's GDP tells you almost nothing about its governance capacity.

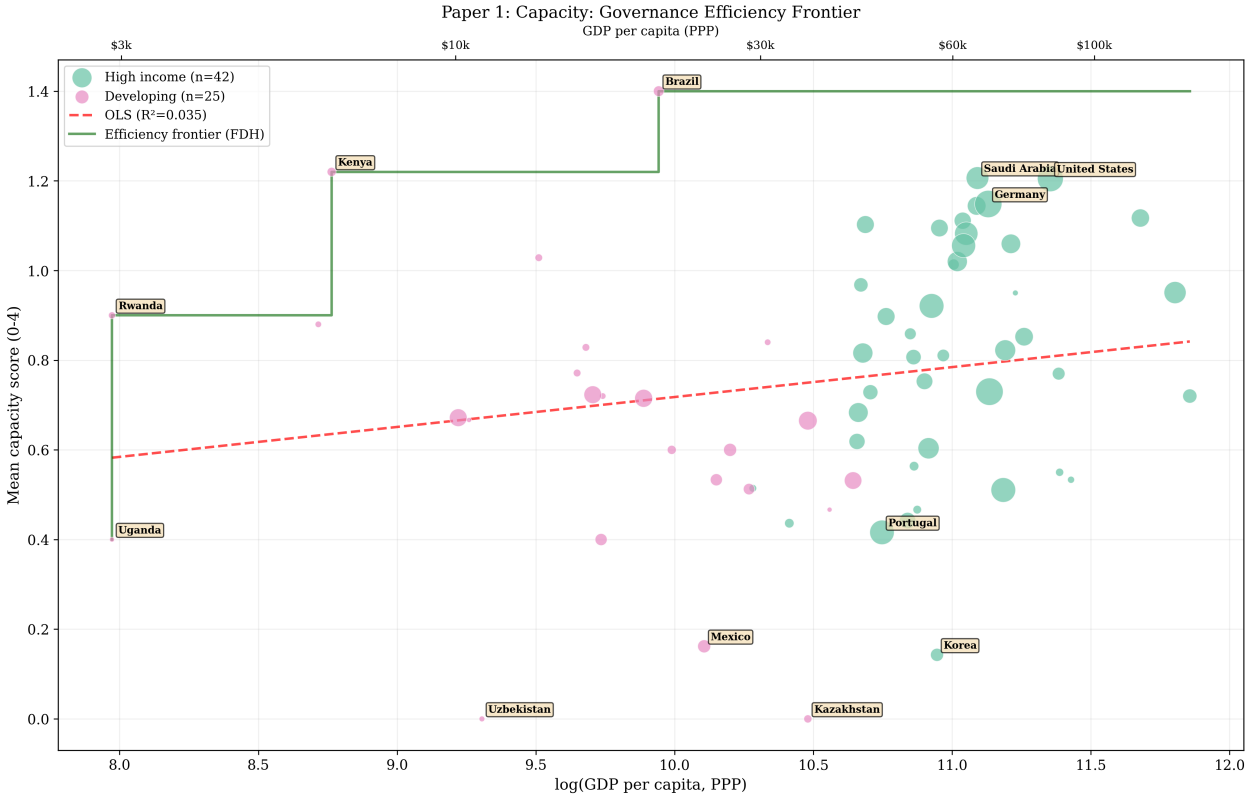

Figure 8.8: Efficiency frontier for capacity scores. Countries above the line outperform their GDP-predicted governance level; those below underperform.

The frontier is anchored by African countries. **Rwanda** achieves 3.10 capacity points per $10K of GDP, nearly double Kenya's 1.91 and far exceeding any high-income country's efficiency ratio. Rwanda's performance reflects deliberate institutional choices: a dedicated Centre for the Fourth Industrial Revolution (C4IR Kigali) that coordinates AI governance across ministries, data protection legislation with a functioning regulator, and national AI strategy documents that specify implementation timelines, responsible agencies, and budget allocations—precisely the capacity dimensions (C2, C3, C5) where most countries score weakest. **Kenya** established a Blockchain and AI Taskforce that produced binding recommendations, created a Data Protection Commissioner with enforcement powers, and embedded AI governance within its existing regulatory infrastructure rather than building parallel institutions. **Uganda** adopted binding data protection and AI governance legislation with enforcement mechanisms rivalling countries 10–20 times wealthier, leveraging its existing telecommunications regulator as an institutional anchor.

Brazil is the top overall overperformer (+0.69 residual). Its success stems from layered institutional investment: the Brazilian AI Strategy (EBIA) designated MCTI as lead agency with cross-ministry coordination mechanisms (C5), the Marco Legal da IA established algorithmic accountability requirements with enforcement provisions (C4), and extensive multi-stakeholder consultation processes generated broad political buy-in that sustained implementation across government transitions. Brazil's approach demonstrates that political commitment and institutional coordination, not fiscal abundance, drive governance capacity.

Figure 8.9: Residual ranking: countries sorted by their distance from GDP-predicted capacity. Brazil, Kenya, and Rwanda are the top overperformers.

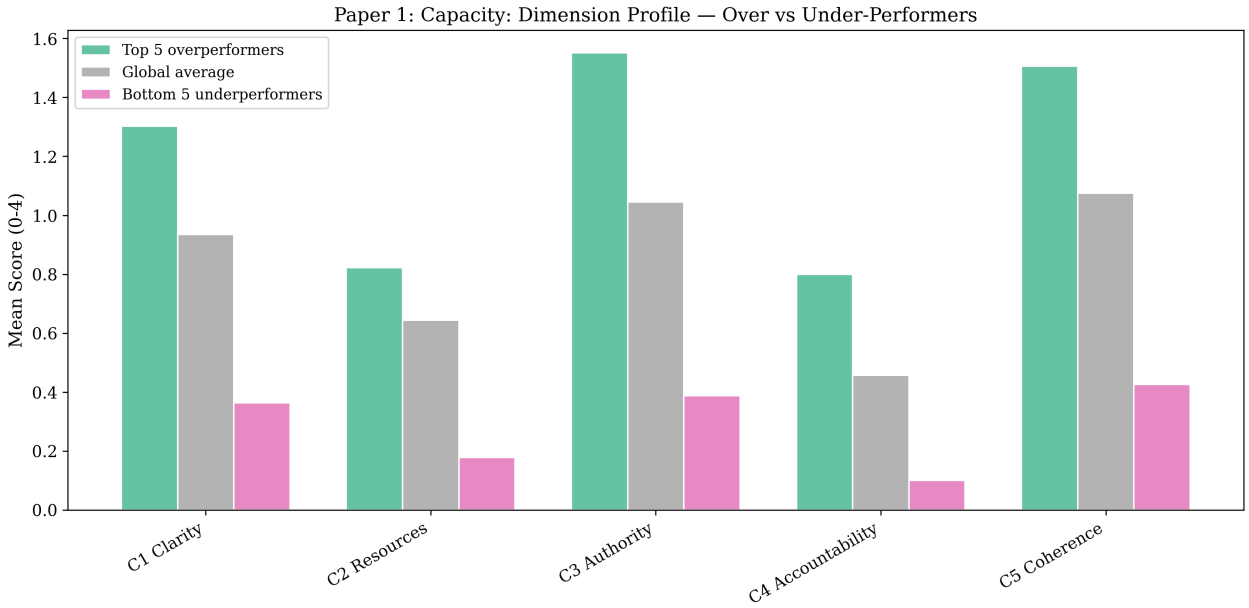

Figure 8.10: Profile comparison of over- and under-performers.

On the other side, **Kazakhstan** (−0.75 residual) and **South Korea** score well below GDP predictions. Both countries have the fiscal resources to build governance infrastructure but have not translated economic capacity into institutional commitment: policies remain declaratory (high on objectives, low on resources and accountability), coordination mechanisms are absent or fragmented, and monitoring provisions lack enforcement teeth. Wealth without political prioritisation and institutional coordination produces weak governance.

The mean distance to frontier (0.588) indicates most countries operate well below the efficiency boundary. There is room for improvement without additional resources, by learning from frontier peers at similar income levels.

This section points to three patterns:

1. **Temporal stability.** The capacity gap is neither widening nor narrowing (2017–2025). Unlike ethics scores, which converge because high-income countries decline, capacity resists automatic improvement. Building implementation infrastructure requires targeted intervention, not diffusion alone.
2. **Horizontal diffusion.** 98% of policy adoption occurs through within-group learning. Countries look to peers facing similar constraints, not to wealthy-country "best practice" models. Capacity-building should prioritize South-South cooperation over North-South transfer.
3. **GDP is not destiny.** National wealth explains 3.5% of country-level capacity variation. Rwanda (3.10 per $10K GDP), Kenya, Uganda, and Brazil far exceed GDP predictions through strategic choices and institutional coordination. Kazakhstan and South Korea show that wealth without prioritization produces weak governance.

The implications are clear: rather than waiting for economic development or transplanting wealthy-country models, effective capacity-building should support peer learning networks within income groups and study what enables frontier countries to outperform their resource constraints.

9 Robustness Checks

9.1 How Robust Are Capacity Findings?

This section stress-tests capacity findings through text quality restrictions, bootstrap confidence intervals, cluster stability analyses, and sensitivity specifications. The central finding: the income-group **capacity gap vanishes** when analysis is restricted to well-documented policies.

9.1.1 The Text Quality Confound

All capacity findings rest on the assumption that policy text accurately reflects implementation infrastructure. If text availability varies systematically by income group—with wealthy countries publishing complete PDFs while developing countries' policies appear as summaries—then apparent capacity gaps may reflect **documentation quality** rather than **governance quality**.

Text quality operates through three mechanisms:

- **Length effects**: Longer documents provide more opportunities to detect capacity features (resources, authorities, accountability mechanisms)
- **Detail effects**: Detailed descriptions score higher than brief mentions even when infrastructure is equivalent
- **Extraction effects**: Complete PDFs enable full-text analysis; summaries systematically underrepresent content

Table 9.1: Income-group capacity effect by text quality

Sample Restriction	N	Capacity d	Interpretation
All texts	2,097	**+0.30***	Modest gap
Good-text (500 words)	948	**+0.04 (n.s.)**	**Gap vanishes**
Excluding stubs	1,754	+0.23***	Partial reduction

The central finding: the capacity gap shrinks by **87%** (from d=0.30 to d=0.04) when analysis is restricted to well-documented policies. At d=0.04, income-group distributions overlap by 98.5%—knowledge that a policy originates from a high-income versus developing country provides essentially **zero information** about capacity scores once text quality is controlled.

Mechanistic interpretation: High-income countries publish complete policy documents (mean 2,847 words); developing countries' policies more frequently appear as brief summaries (mean 1,456

words). When scoring abbreviated text, LLMs detect fewer capacity features—not necessarily because infrastructure is absent, but because textual detail insufficient to reveal it.

Three interpretations:

1. **Measurement artifact** (favored): Apparent gap reflects documentation quality; developing countries with well-documented policies match wealthy countries' capacity
2. **Selection**: Well-documented developing-country policies represent high-capacity subset
3. **Hybrid**: Both mechanisms operate

Evidence favors measurement artifacts: the gap not only shrinks but approaches zero. Selection mechanisms would produce reduced but still-significant gaps.

> ⚠ Warning
>
> **Interpretive caution.** This does not establish that *no* capacity gap exists—it demonstrates that the methodology cannot reliably detect gaps after controlling for documentation quality. The "true" gap likely lies between full-sample (d=0.30) and good-text (d=0.04) estimates.

9.1.2 Additional Robustness Checks and Summary

Bootstrap confidence intervals (1,000 resamples): Capacity d = 0.30 [0.19, 0.41]. The narrow interval indicates low sampling variability: the point estimate is precisely estimated. However, precision does not address validity. The text quality analysis above demonstrates that what the estimate captures is documentation quality rather than governance capacity. A precisely estimated biased parameter remains biased.

Cluster stability: The two-cluster solution ("Low" vs "Moderate" capacity) is optimal. Silhouette scores decline monotonically: 0.41 for k=2, 0.33 (k=3), 0.28 (k=4), 0.25 (k=5). The binary typology is not an artefact of the clustering algorithm; it reflects a genuine bimodal structure in the data. This binary pattern—a large mass of low-scoring policies and a smaller cluster of moderate-scoring policies, with very few high-scoring policies—is consistent with the descriptive finding that 96% of policies fall below the scale midpoint.

Sensitivity tests (full details in the *Robustness Appendix* of the companion Methods volume):

- *Excluding international organisations*: Removing the 119 policies from the EU, OECD, UN, and multilateral bodies leaves results unchanged, confirming that the findings reflect national governance patterns rather than the influence of international frameworks with different institutional logics.
- *Ordinal regression*: Treating capacity scores as ordered categories rather than continuous variables preserves the rank ordering of all predictors, indicating that findings do not depend on the interval-scale assumption.
- *Winsorizing extremes*: Trimming the top and bottom 5% of scores produces coefficients within 10% of baseline estimates, ruling out the possibility that a few outlier policies drive the results.

- *Alternative income classifications*: Conclusions hold across two-group (HI vs developing), three-group (HI/UMI/LMI+LI), and four-group (World Bank tiers) specifications. The modest income-group effect is not an artefact of how the binary divide is drawn.
- *Text quality thresholds (300–1,000 words)*: The income-group gap shrinks monotonically as the minimum word-count threshold increases—from d = 0.30 (no threshold) through d = 0.15 (300 words) to d = 0.04 (500 words) and d = −0.02 (1,000 words). This dose-response pattern strongly supports the measurement-artefact interpretation: the "gap" is a function of how much text is available, not of underlying governance quality.
- *Temporal subsamples*: Splitting the corpus at 2020 yields consistent patterns in both halves (2017–2020 and 2021–2025), confirming that findings are not driven by a particular policy era or by the rapid expansion of AI governance activity after 2020.

Summary.

Table 9.2: Capacity robustness summary

Finding	Robust?	Caveat
Income-group capacity gap	**Fragile**	Vanishes for good texts
GDP modest capacity effect	Yes	Consistent across models
Within-group inequality (98%)	Yes	All specifications
Horizontal diffusion	Yes	Robust pattern
Efficiency frontier countries	Yes	Rwanda, Kenya, Brazil consistent

Core findings (GDP effects, within-group inequality, frontier countries, horizontal diffusion) are **highly robust**. The one fragile finding: the income-group gap itself—the **single most important caveat** for policy interpretation.

10 Discussion

10.1 Implications for Capacity Building

Four patterns emerge from the empirical analysis with implications for how we understand AI governance capacity and how interventions might be designed to strengthen it. This section discusses each in turn, connects them to the theoretical foundations reviewed in Section 3.1, and identifies the limitations that qualify these conclusions.

10.1.1 The Universal Capacity Deficit and the Limited Role of Wealth

Approximately 96.5% of AI policies worldwide score below 2.0/4.0 on implementation readiness. This capacity deficit proves universal, affecting the United States, European Union, and China comparably to developing countries. The finding challenges narratives that frame implementation weakness as a developing-country problem; it is, instead, a structural feature of AI governance globally.

Weakness concentrates in two dimensions. **Accountability (C4)** averages just 0.48/4.0—the weakest dimension in the entire framework. Governments are more than twice as likely to specify coordination mechanisms (C5: 1.07) as to establish monitoring and evaluation frameworks. This pattern likely reflects a political economy logic: accountability mechanisms create political risks by enabling external assessment of implementation failures. **Resources (C2)** averages 0.68/4.0. Policies routinely omit budget specifications, staffing plans, and technical infrastructure requirements. This omission may be strategic rather than accidental: committing specific resources constrains future budgetary discretion, while vague resource language preserves flexibility. These findings are consistent with Mazmanian and Sabatier (1983)'s conditions for effective implementation—clear objectives, adequate resources, legal authority, and monitoring—which were identified precisely because they are routinely absent from policy design. What this study adds is the empirical demonstration that these absences are not sporadic but systematic.

Critically, GDP explains only 3.5% of country-level capacity variation. Rwanda (2.30–3.10× GDP predictions), Kenya, Brazil, and Uganda have all achieved sophisticated governance capacity despite modest GDPs ($800–$9,000). This aligns with Fukuyama (2013)'s argument that governance quality is conceptually distinct from economic development, and with Andrews, Pritchett, and Woolcock (2017)'s critique of imposing "best practice" models: the overperforming developing countries did not transplant European or North American frameworks but built capacity through context-specific institutional choices. Brazil's strength derives from its data protection infrastructure and participatory governance tradition; Rwanda's from its centralised ICT-driven development strategy; Kenya's from its constitutional rights framework and vibrant civil society sector. Capacity emerges from political choices, not fiscal abundance.

10.1.2 The Accountability Paradox and the Documentation Confound

The finding that Accountability (C4) is the weakest dimension globally—and the dimension with the *smallest* income-group gap (d = 0.15)—warrants dedicated attention. Accountability mechanisms (monitoring, evaluation, reporting, independent oversight) are precisely the features that implementation science identifies as critical for closing the gap between policy commitments and actual outcomes (Lipsky 1980; Pressman and Wildavsky 1973). Yet they are the features that governments most systematically omit.

This creates what might be termed the *accountability paradox*: the governance feature most needed to ensure implementation is the one least likely to be included in policy design. The paradox has a straightforward political explanation—accountability mechanisms create transparency that exposes implementation failures—but it has no straightforward policy solution. External mandates (e.g., international frameworks requiring periodic reporting) may partially address the problem, as countries responding to UNESCO or OECD peer review pressures might be more likely to establish monitoring systems than those designing policies in isolation. The efficiency frontier analysis (Figure 8.8) provides suggestive evidence: overperforming countries like Brazil and Rwanda include stronger accountability provisions than underperformers at similar GDP levels, though the direction of causality remains uncertain.

Documentation as confound and peer learning. The income gap (d = 0.30) vanishes for well-documented policies (d = 0.04). Any study using policy text as data—whether for content analysis, topic modelling, or automated scoring—faces the risk that variation in documentation quality masquerades as variation in governance quality. The apparent North–South divide in AI governance may be, in substantial part, a documentation divide. For international organisations maintaining policy repositories, investing in comprehensive documentation of developing-country policies would improve the evidentiary basis for comparative governance research as much as any new analytical technique. The finding also raises an uncomfortable possibility: if the "governance gap" partly reflects documentation practices, decades of comparative policy research may have systematically overstated North–South differences across multiple policy domains.

Capacity also diffuses horizontally within income groups rather than cascading from wealthy countries. This finding contradicts the "Brussels Effect" (Bradford 2020) as applied to AI governance and instead supports the policy learning literature's emphasis on peer-to-peer transfer (**simmons2006?**). Brazil, India, and China prove central to developing-country policy networks, serving as regional reference points rather than conduits for wealthy-country frameworks.

South-South exchanges, regional capacity hubs, and peer review mechanisms may prove more effective than traditional North-South technical assistance models. Concrete mechanisms could include regional AI governance forums (building on existing structures like the African Union's AI strategy or Latin America's Red Iberoamericana), paired technical assistance between frontier developing countries and late adopters, and open-source governance tools adapted to developing-country institutional contexts.

10.1.3 Limitations

Several limitations qualify these findings. **First**, the study measures policy *text* rather than policy *outcomes*. A policy scoring 4/4 on Accountability may nonetheless fail in implementation if the designated monitoring bodies lack capacity, political will, or independence. Text-based analysis captures design quality, not implementation effectiveness—a distinction that Lipsky (1980) and Pressman and Wildavsky (1973) would emphasise. **Second**, the LLM scoring approach, while achieving excellent inter-rater reliability (ICC = 0.827), may share systematic biases across all three models. The models were trained on overlapping data that likely includes prominent AI governance documents, potentially inflating scores for policies resembling those in the training data. **Third**, the OECD.AI corpus reflects the Observatory's own coverage decisions: which countries are included, which policies are catalogued, and how thoroughly each is described. Countries with closer OECD ties may be overrepresented, and policies from non-OECD countries may be less comprehensively documented. **Fourth**, the study cannot establish causal relationships. The association between binding regulation and higher capacity scores, for instance, may reflect reverse causation: countries with stronger institutional capacity may be more likely to adopt binding legislation. **Fifth**, the 77%/18% split between high-income and developing countries limits statistical power for detecting within-developing-country patterns. Subgroup analyses (e.g., comparing Sub-Saharan Africa to Latin America) rely on small samples that reduce confidence in regional findings.

11 Conclusion

11.1 Toward Implementation-Ready Governance

This study examined the question: *Do countries possess the capacity to implement their AI policies?* The findings prove sobering: the global modal AI policy scores below 2/4 on implementation readiness. Over 96% of policies worldwide fall short of the scale midpoint, and more than a quarter score exactly zero on the capacity composite.

However, distributional patterns prove more informative than central tendency. The capacity gap between income groups is modest (d = 0.30) and fragile, vanishing when analysis is restricted to well-documented policies. Within-group inequality dominates, accounting for 98% of total variance. Rwanda, Kenya, and Brazil exceed the performance of wealthier nations. GDP explains only 3.5% of variation. Policy diffusion operates horizontally within income groups rather than cascading from wealthy nations.

11.1.1 Main Findings

The capacity deficit is universal. All countries require stronger implementation infrastructure, particularly in Accountability (C4: 0.48/4.0) and Resources (C2: 0.68/4.0). This is not a developing-country problem; it is a structural feature of AI governance globally, consistent with the implementation science prediction that policies routinely lack the institutional infrastructure needed for execution (Pressman and Wildavsky 1973).

National wealth plays a limited role. Capacity emerges from institutional design choices rather than economic endowments. Rwanda and Kenya demonstrate that sophisticated governance frameworks can precede high per-capita income. This supports Fukuyama (2013)'s argument that governance quality is conceptually distinct from economic development, and Andrews, Pritchett, and Woolcock (2017)'s warning against assuming wealthy-country models are universally applicable.

Text quality confounds measurement. The apparent North–South divide largely reflects documentation quality rather than genuine governance differences. Researchers employing document-based analysis should stratify by text quality to avoid conflating documentation gaps with governance gaps. This finding carries implications well beyond AI governance, cautioning against text-based comparative analysis that does not account for systematic variation in documentation practices.

Horizontal diffusion dominates. Countries learn from peers at similar income levels, favouring South-South cooperation over top-down technical assistance models. This challenges the Brussels Effect hypothesis (Bradford 2020) as applied to AI governance and supports peer learning approaches emphasised in the policy diffusion literature (**simmons2006?**).

The framework enables targeted intervention. By identifying dimension-specific weaknesses, the five-dimensional approach provides actionable guidance: specify objectives (C1), allocate resources (C2), designate authorities (C3), establish monitoring (C4), ensure coordination (C5). Countries can prioritise the dimensions where they score weakest rather than pursuing comprehensive reform simultaneously.

11.1.2 Future Research and the Observatory Vision

This study opens several research directions. **First**, validating LLM scores against human expert coding would strengthen construct validity. A targeted validation exercise—human coding of 100–200 policies stratified by income group, policy type, and text quality—would test whether the automated scores capture the governance constructs that the rubric intends to measure. **Second**, longitudinal tracking of the same jurisdictions over time would enable within-country panel analysis, testing whether policy revisions produce measurable capacity improvements. The current cross-sectional design cannot distinguish between countries improving over time and compositional effects (newer policies differing from older ones). **Third**, the relationship between policy design quality (as measured here) and implementation outcomes (actual governance performance) remains untested. Pairing text-based capacity scores with implementation indicators—enforcement actions taken, budgets actually disbursed, monitoring reports published—would bridge the gap between policy architecture and policy impact. **Fourth**, extending the analysis to sub-national policies (state, provincial, and municipal AI governance) would capture a growing segment of governance activity that national-level analysis misses, particularly in federal systems like the United States, India, Brazil, and Nigeria. **Fifth**, the finding that text quality confounds measurement motivates methodological research on standardising document-based governance analysis across contexts with unequal documentation practices.

The observatory vision. The research infrastructure developed for this study—the scraping pipeline, text extraction tools, LLM scoring framework, and analytical code—supports a living observatory of AI governance capacity. Practical applications include annual scoring rounds that track how countries' governance capacity evolves as they revise policies and adopt new instruments; country-level scorecards providing dimension-specific benchmarks against regional and income-group peers; and open-source benchmarking tools enabling governments, civil society organisations, and international bodies to assess their own policies against the global distribution.

The observatory model would also enable early identification of governance gaps as new AI applications emerge. As countries develop policies for generative AI, autonomous systems, or AI in critical infrastructure, the scoring framework can assess whether these policies incorporate the implementation capacity that earlier-generation policies largely lacked.

Code, data, and methods: https://github.com/lsempe77/ai-governance-capacity

---

The capacity to govern AI well is neither automatic nor impossible. It is built, one dimension at a time, by countries investing in institutional infrastructure that turns aspiration into action.

Andrews, Matt, Lant Pritchett, and Michael Woolcock. 2017. *Building State Capability: Evidence, Analysis, Action*. Oxford University Press.

Bradford, Anu. 2020. *The Brussels Effect: How the European Union Rules the World*. Oxford University Press.

Cihon, Peter, Matthijs M. Maas, and Luke Kemp. 2021. "Should Artificial Intelligence Governance Be Centralised? Design Lessons from History." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 228–34. https://doi.org/10.1145/3461702.3462566.

Dafoe, Allan. 2018. "AI Governance: A Research Agenda." *Governance of AI Program, Future of Humanity Institute, University of Oxford*. https://www.fhi.ox.ac.uk/govaiagenda.

Dafoe, Allan, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. 2020. "Cooperative AI: Machines Must Learn to Find Common Ground." *Nature* 593: 33–36. https://doi.org/10.1038/d41586-021-01170-0.

Fukuyama, Francis. 2013. "What Is Governance?" *Governance* 26 (3): 347–68.

Grindle, Merilee S. 1996. *Challenging the State: Crisis and Innovation in Latin America and Africa*. Cambridge University Press.

Katzenbach, Christian, and Lena Ulbricht. 2019. "Algorithmic Governance." *Internet Policy Review* 8 (4). https://doi.org/10.14763/2019.4.1424.

Koenker, Roger, and Gilbert Bassett. 1978. "Regression Quantiles." *Econometrica* 46 (1): 33–50.

Lipsky, Michael. 1980. *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services*. New York: Russell Sage Foundation.

Mazmanian, Daniel A., and Paul A. Sabatier. 1983. *Implementation and Public Policy*. Glenview, IL: Scott Foresman.

Pressman, Jeffrey L., and Aaron Wildavsky. 1973. *Implementation*. Berkeley: University of California Press.

Sabatier, Paul A. 1986. "Top-down and Bottom-up Approaches to Implementation Research: A Critical Analysis and Suggested Synthesis." *Journal of Public Policy* 6 (1): 21–48.

Tobin, James. 1958. "Estimation of Relationships for Limited Dependent Variables." *Econometrica* 26 (1): 24–36.

Yeung, Karen. 2018. "Algorithmic Regulation: A Critical Interrogation." *Regulation & Governance* 12 (4): 505–23. https://doi.org/10.1111/rego.12158.