Global Observatory of AI Governance Capacity

Measuring Implementation Readiness Across 2,100+ Policies

Lucas Sempé

February 11, 2026

Table of contents

1 Global Observatory of AI Governance Capacity

Measuring Implementation Readiness Across 2,100+ Policies

Preface

This book presents the first systematic global assessment of AI governance **capacity** — measuring whether governments possess the institutional infrastructure to implement their AI policies.

Drawing on 2,100+ policies across 70+ jurisdictions, we score each policy on five dimensions of implementation capacity: clarity, resources, authority, accountability, and coherence. The analysis reveals that implementation readiness varies dramatically—not primarily between income groups, but within them.

1.1 Key Findings

- **Text quality confound**: Income-group gaps vanish when restricted to well-documented policies
- **Within-group inequality**: 98% of variation occurs within income groups rather than between them
- **Efficiency frontier**: Brazil, Kenya, Rwanda, Tunisia outperform GDP predictions
- **Horizontal diffusion**: Countries learn from income-group peers, not top-down cascades

1.2 Methodology

We employ an LLM ensemble (Claude Sonnet 4, GPT-4o, Gemini Flash 2.0) achieving ICC = 0.827 (excellent inter-rater reliability), enabling policy analysis at unprecedented scale.

Citation: Sempé, L. (2026). *Global Observatory of AI Governance Capacity.* 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

> **Chapter summary.** This chapter motivates the central research question: Do governments have the institutional capacity to implement their AI policies? We identify gaps in the literature, preview the analytical framework, and summarize key findings.

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. Yet this proliferation raises a fundamental question that governance research has largely neglected:

Can governments actually implement their AI policies?

Implementation capacity—the institutional infrastructure enabling policy execution—remains unmeasured in AI governance. We know what policies say but not whether governments possess the clarity, resources, authority, accountability structures, and coordination mechanisms to translate commitments into action.

2.1.2 What We Measure

This book develops and applies a comprehensive capacity assessment framework across five dimensions:

1. **C1 Clarity**: Specificity of policy objectives and implementation plans
2. **C2 Resources**: Budgets, staffing, technical infrastructure
3. **C3 Authority**: Legal mandates, enforcement powers, regulatory competence
4. **C4 Accountability**: Oversight mechanisms, reporting requirements, review processes
5. **C5 Coherence**: Inter-agency coordination, consistency with existing frameworks

Each dimension is scored 0-4 based on detailed rubrics grounded in implementation science (Mazmanian and Sabatier 1983; Pressman and Wildavsky 1973).

2.1.3 Key Findings Preview

Our analysis of 2,100+ policies reveals:

- **Low average capacity**: Mean 1.38/4.0, indicating most policies lack robust implementation infrastructure
- **Text quality confound**: Income-group gaps (d=0.30) vanish (d=0.04) for well-documented policies
- **Within-group dominance**: 98% of variation occurs within income groups, not between them
- **Efficiency frontier**: Brazil, Kenya, Rwanda, Tunisia outperform GDP-predicted levels
- **Horizontal diffusion**: Countries learn from income-group peers rather than following wealthy-country templates

2.1.4 Roadmap

- **Chapter 5** maps the global landscape of governance capacity
- **Chapter 6** examines socioeconomic and institutional determinants
- **Chapter 7** analyzes inequality patterns
- **Chapter 8** traces temporal dynamics and policy diffusion
- **Chapter 14** presents robustness checks, especially the text quality confound
- **Chapters 15-16** discuss implications and conclusions

The following chapters reveal a capacity landscape more complex than simple North-South divides suggest.

3 Literature Review

3.1 Theoretical Foundations

> **Chapter summary.** We ground AI governance capacity measurement in implementation science and state capacity theory, showing how classic frameworks from Mazmanian, Sabatier, Lipsky, and Grindle map onto contemporary AI governance challenges.

3.1.1 Implementation Science and Policy Capacity

The study of policy implementation, beginning with Pressman and Wildavsky's (1973) observation that well-designed programs routinely fail execution, provides the conceptual foundation for assessing AI governance capacity. Implementation science asks: what features separate policies that can be implemented from those remaining aspirational?

3.1.1.1 Top-Down Approaches

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 map directly onto our capacity dimensions:

- **Clarity**: Clear objectives and causal theories
- **Resources**: Committed and skilled officials

- **Authority**: Legal structuring and enforcement powers
- **Accountability**: Monitoring and oversight mechanisms
- **Coherence**: Coordination across implementing agencies

Sabatier (1986) synthesized top-down and bottom-up perspectives, arguing that both formal structure and implementing strategies matter. This informs our attention to policy architecture—the institutional infrastructure enabling implementation.

3.1.1.2 Bottom-Up Approaches and Discretion

Lipsky (1980) shifted focus to "street-level bureaucrats"—frontline workers whose discretionary decisions effectively make policy. In AI governance, this insight proves critical: even comprehensive legislation fails if regulators lack expertise, resources, or mandate to exercise oversight.

Data protection authorities interpreting GDPR's algorithmic accountability provisions, or competition regulators assessing AI market power, exemplify street-level discretion. Our **Accountability (C4)** dimension captures whether policies constrain this through monitoring and evaluation frameworks.

3.1.1.3 State Capacity

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

Fukuyama (2013) argued that governance quality is conceptually distinct from democracy or GDP, proposing measurement through government outputs. Our scoring framework operationalizes this—measuring institutional readiness through policy quality rather than inputs like national wealth.

Andrews, Pritchett, and Woolcock (2017) introduced "building state capability" through iterative adaptation, arguing against imposing "best practice" from high-income countries. This resonates with our finding that developing countries achieve governance through different pathways than wealthy nations.

3.1.2 The Capacity Gap in Digital Governance

Recent research documents persistent implementation gaps in digital regulation. (**yeung2018?**) shows algorithmic regulation demands technical expertise most governments lack. (**katzenbach2019?**) demonstrates that platform governance creates enforcement challenges traditional regulators struggle to address.

(**dafoe2018?**) 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.

3.1.3 Measurement Challenges

Despite theoretical progress, **systematic capacity measurement remains rare**. Existing studies rely on qualitative assessments (**cihon2021?**) or expert surveys (**dafoe2020?**) that don't scale. Our LLM-based scoring methodology addresses this gap, enabling comprehensive assessment across 2,100+ policies.

3.1.4 Contribution

This book operationalizes classical implementation science for AI governance capacity measurement. We provide:

1. **Validated measurement framework**: Five capacity dimensions grounded in Mazmanian and Sabatier (1983), Lipsky (1980), and Grindle (1996)
2. **Scalable methodology**: LLM ensemble achieving ICC = 0.827 (excellent reliability)
3. **Global dataset**: First comprehensive capacity assessment across 70+ jurisdictions
4. **Empirical findings**: Testing whether wealth determines governance capacity

The following chapters reveal that implementation readiness varies more within income groups than between them, challenging assumptions about governance divides.

4 Data & Methods

4.1 The OECD.AI Corpus

Chapter summary. This chapter describes the data collection pipeline: from the OECD.AI Policy Observatory through document retrieval, text extraction, and quality classification. We detail the construction of a 2,216-policy corpus with 11.4 million words of analysis-ready text across 70+ jurisdictions.

4.1.1 Data Source

Our data come from the **OECD.AI Policy Observatory** (OECD 2024), the most comprehensive international tracker of AI policy initiatives. Established as a collaborative effort among OECD member states and partner countries, the Observatory serves as the global standard for monitoring AI governance activity. It catalogues government actions related to AI — including national strategies, legislation, executive orders, guidelines, and programmes — with structured metadata on jurisdiction, year, policy type, target sectors, and responsible organisations. This structured approach makes the Observatory uniquely suited for systematic cross-national comparison, as each entry follows a consistent documentation schema that enables quantitative analysis at scale.

We politely scraped the complete Observatory as of January 2026, obtaining **2,216 policy entries** spanning **70+ jurisdictions** and the years **2017–2025**. This snapshot represents the state of global AI governance at a critical juncture, as many jurisdictions transition from voluntary guidelines to binding regulation.

Table 4.1: Corpus overview

Metric	Value
Total policy entries	2,216
Unique jurisdictions	70+
Time span	2017–2025
Policy types	Strategies, laws, guidelines, executive orders, programmes
Source	OECD.AI Policy Observatory

Table 4.1 shows the breadth of our corpus, which encompasses nearly every documented AI governance initiative globally over the past eight years. The 70+ jurisdictions include not only major

economies but also developing countries in Africa, Asia, and Latin America, providing the geographic diversity necessary to examine capacity gaps across income levels.

4.1.2 Document Retrieval

The OECD.AI Observatory provides brief descriptions (typically <500 words) and links to source documents, but does not host full texts. This design reflects the Observatory's role as a catalog rather than an archive — it points to official documents but leaves them at their original locations. For our analysis, however, we required the complete policy texts to enable detailed assessment of implementation capacity. This necessitated building a retrieval pipeline capable of locating and downloading documents that might have moved, been renamed, or disappeared from their original URLs.

Our five-strategy retrieval pipeline operated as a cascading fallback system. First, we attempted direct downloads from the `source_url` field provided in the Observatory metadata, which succeeded for approximately 60% of entries. For documents where direct download failed, we scraped the OECD.AI web page for each policy entry to locate embedded source links that might not appear in the structured metadata. When original URLs had moved or expired — a common occurrence for policy documents published years earlier — we queried the Internet Archive Wayback Machine to retrieve historical snapshots. For documents unavailable through any of these channels, we conducted targeted searches using DuckDuckGo with carefully constructed queries combining the policy title, jurisdiction, and file type restrictions. Finally, for the most difficult cases, we employed the Claude API's web search capability to locate official document URLs through more sophisticated reasoning about likely hosting locations.

This layered approach achieved approximately 94% coverage, successfully retrieving around 2,085 documents to local storage. The remaining entries — primarily press releases, brief announcements, or policies documented only through secondary sources — remained available as OECD snippets, providing at least minimal text for analysis even when full documents proved inaccessible.

4.1.3 Text Extraction

Retrieving documents was only the first challenge; extracting clean, analysis-ready text from diverse file formats proved equally demanding. Policy documents arrive in varied formats — PDFs may be text-based or scanned images, web pages may embed content within complex navigation structures, and documents may span from single-page executive summaries to hundred-page legislative texts. Each format required specialized handling to extract content accurately while removing headers, footers, page numbers, and other non-substantive elements that would interfere with analysis.

We developed format-specific extraction pipelines matched to document characteristics. For PDF documents — the most common format in our corpus — we employed PyMuPDF (`fitz`), which excels at extracting text from text-based PDFs while preserving document structure. For HTML documents, we used `trafilatura`, a content extraction library specifically designed to identify main textual content while stripping navigation menus, sidebars, and other boilerplate elements typical of government websites. For entries where no downloadable source could be located, we fell back

to the OECD snippet text, accepting the limitation of abbreviated content rather than excluding these policies entirely.

Each document was then classified into one of three quality tiers based on extracted word count, providing a systematic approach to assessing text adequacy for detailed analysis:

Table 4.2: Text quality distribution

Quality Tier	Word Count	N	%	Description
Good	500 words	948	42.8%	Full analysis possible
Thin	100–499 words	806	36.4%	Usable with caveats
Stub	<100 words	462	20.8%	Minimal text only
	Analysis-ready	**1,754**	**79.2%**	Good + Thin

Table 4.2 reveals that nearly 80% of our corpus (1,754 documents) contains sufficient text for reliable analysis, with 43% classified as "Good" quality with substantial content exceeding 500 words. The 806 "Thin" documents — containing 100–499 words — provide enough context for basic scoring but may lack the detail needed to assess more nuanced implementation features. The 462 "Stub" entries, containing fewer than 100 words, typically represent brief announcements or press releases that offer minimal substantive content. While we include these in corpus statistics, they contribute little to the analytical results. The total extracted corpus contains 11.4 million words, with a median document length of 1,247 words (IQR: 318–4,892), indicating that a typical AI governance policy provides several pages of substantive content suitable for detailed assessment.

4.1.4 Enriched Corpus

The retrieval and extraction pipeline produced a unified corpus file (`corpus_enriched.json`) that merges OECD metadata with our extracted content and quality assessments. For each of the 2,216 entries, this file preserves the original OECD metadata — including title, jurisdiction, year, URL, policy type, and target sectors — while adding the extracted full text (or OECD snippet where full text was unavailable), text quality classification, word count, and extraction method employed. This enriched structure enables analyses that link policy content to contextual metadata, supporting questions about how governance quality varies by jurisdiction, year, or policy type.

4.1.5 Country Metadata

To enable cross-national comparison, each jurisdiction was mapped to standardized contextual metadata using World Bank classifications. Income groups follow the World Bank's four-tier system: High Income (HI), Upper Middle Income (UMI), Lower Middle Income (LMI), and Low Income (LI). For analyses focused on the North–South divide, we constructed a binary classification contrasting High Income countries against Developing countries (aggregating UMI, LMI, and LI). Regional classifications employ the World Bank's geographic taxonomy: East Asia & Pacific (EAP), Europe & Central Asia (ECA), Latin America & Caribbean (LAC), Middle East & North Africa

(MENA), North America (NAM), South Asia (SA), and Sub-Saharan Africa (SSA). We also incorporated GDP per capita (current US dollars, 2023) as a continuous measure of economic development, enabling analyses that examine governance quality relative to national wealth.

International organisations — including the OECD itself, the European Union, the United Nations, and multilateral development banks — were flagged separately and excluded from country-level analyses where appropriate, as these entities operate under different institutional logics than national governments.

4.1.6 Sample Composition

The final analytical sample reflects the OECD.AI Observatory's coverage, which skews toward high-income countries:

Table 4.3: Sample by income group

Income Group	N Policies	%	N Countries
High Income	1,700	76.7%	~40
Developing	397	17.9%	~30
International	119	5.4%	—
Total	**2,216**	**100%**	**70+**

Table 4.3 reveals a substantial compositional imbalance: high-income countries account for 77% of policies in the corpus, while developing countries contribute only 18%. This disparity reflects the genuine distribution of AI governance activity globally — high-income countries have produced more policies, published more documentation, and maintained more accessible policy archives. However, this imbalance creates analytical challenges, as conventional statistical comparisons assume relatively balanced groups. We address potential selection effects and the implications of unbalanced samples through comprehensive robustness checks in Section 10.1, including analyses restricted to well-documented policies and country-level aggregations that equalize representation.

4.1.7 Analytical Pipeline Overview

The journey from raw OECD.AI metadata to empirical findings involves multiple transformation stages, each addressing distinct methodological challenges. Figure 4.1 visualizes this progression, showing how 2,216 initial entries flow through retrieval, extraction, scoring, and analysis to produce the 120 outputs (figures, tables, statistical tests) that appear in subsequent chapters. This pipeline architecture separates data collection concerns from analytical decisions, enabling transparent documentation of how each methodological choice affects downstream results.

Figure 4.1 shows how each stage transforms the data: from initial policy entries through document retrieval and text extraction (the data collection phase documented in preceding sections), to LLM-based scoring (detailed in Section 5.1), culminating in the 20 analytical chapters that follow. The 6,641 LLM API calls represent three model assessments for each of the 2,216 policies across 10 dimensions, with the ensemble approach ensuring reliability through inter-model agreement.

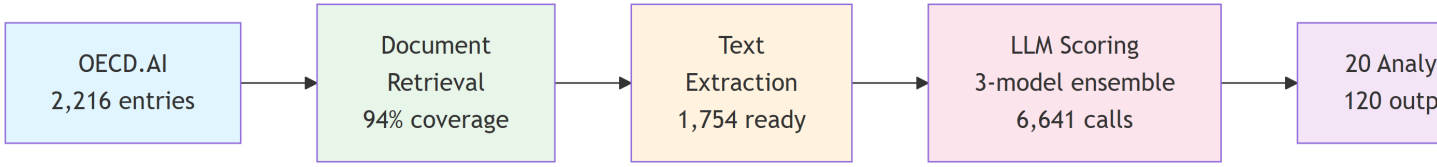

Figure 4.1: Analytical pipeline from corpus to results

4.1.8 Analytical Methods

The statistical analyses in subsequent chapters employ multiple complementary methods to examine governance capacity from different angles. This methodological pluralism enables robust inference: findings that emerge consistently across diverse analytical approaches inspire greater confidence than those dependent on a single modeling choice. Here we overview the core analytical techniques; specific model specifications appear in their respective chapters.

4.1.8.1 Text-to-Data Conversion: LLM Ensemble Scoring

The foundational methodological step — and the innovation that enables analysis at this scale — is the conversion of unstructured policy documents into structured quantitative scores. Unlike traditional text analysis approaches that extract word frequencies, topics, or sentiment, our method employs frontier large language models as expert policy analysts. Each LLM reads the full policy document (up to the model's context window, typically 8,000+ words), applies the detailed scoring rubric for all 10 dimensions simultaneously, and returns structured JSON-formatted scores with textual evidence justifying each assessment. This approach preserves the interpretive sophistication of human expert coding — capturing whether a policy merely mentions implementation features or provides concrete operational details — while achieving the scale necessary to analyze 2,216 documents.

The three-model ensemble (Claude Sonnet 4, GPT-4o, Gemini Flash 2.0) functions as a panel of expert raters, with the median score serving as the final assessment. This ensemble design addresses the known variability of individual LLM outputs while leveraging their complementary strengths: Claude's nuanced policy interpretation, GPT-4o's balanced analytical approach, and Gemini's efficient processing. The resulting ICC(2,1) = 0.827 demonstrates excellent inter-rater reliability, comparable to or exceeding typical human coder agreement on complex policy dimensions. Detailed validation of this approach, including comparison with human expert ratings, appears in Section 5.1. All subsequent statistical analyses operate on these LLM-derived scores rather than on raw text, treating the scoring outputs as the primary data.

4.1.8.2 Descriptive Analysis

Each analytical chapter begins with descriptive statistics and visual exploration. We present dimension-specific distributions using histogaps (histograms with frequency annotations), ridge plots showing density distributions across groups, and radar charts illustrating multidimensional profiles.

These visualizations reveal patterns that summary statistics alone might obscure — such as bimodality in score distributions or dimension-specific gaps that disappear in composite scores. Box plots with violin overlays show both central tendency and full distributional shape, while heatmaps reveal clustering patterns in policy portfolios across countries and dimensions.

4.1.8.3 Regression Models

Chapters examining determinants of governance capacity employ four complementary regression approaches. Standard OLS regression establishes baseline relationships between predictors (GDP per capita, policy year, document type, text quality) and capacity scores. Multilevel models with random intercepts for countries account for the nested structure of policies within jurisdictions, correcting for dependency that would otherwise inflate standard errors. Quantile regression examines whether predictors affect low-scoring and high-scoring policies differently, revealing heterogeneous effects across the distribution. Tobit models address the substantial floor effect (27.6% of policies score exactly zero) through left-censoring at zero, correcting the attenuation bias that OLS exhibits when floor effects are present.

4.1.8.4 Inequality Analysis

The inequality chapters employ decomposition techniques to partition total variance into meaningful components. Gini coefficients and Lorenz curves quantify overall inequality in governance scores and visualize concentration. Theil's T index enables exact additive decomposition of total inequality into between-group (high-income vs. developing) and within-group components, revealing how much of the apparent North–South divide reflects genuine group differences versus within-group heterogeneity. Policy portfolio analysis examines breadth (whether countries address all dimensions) versus depth (score levels within covered dimensions), distinguishing coverage gaps from implementation quality.

4.1.8.5 Temporal Analysis

Chapters examining governance dynamics over time use panel data methods to separate within-country trends from between-country differences. First-difference models examine year-to-year changes, removing country fixed effects to focus on temporal evolution. We employ Cohen's d effect sizes to assess the substantive significance of changes over time, complementing statistical significance tests that can be misleading with large samples. Convergence analysis tests whether the gap between income groups is narrowing, widening, or remaining stable, using interaction terms between income group and time trends.

4.1.8.6 Multivariate Methods

Principal component analysis (PCA) examines the latent structure underlying the 10 governance dimensions, testing whether capacity and ethics represent empirically distinct constructs. We report eigenvalues, scree plots, and component loadings to assess dimensionality, applying the Kaiser

criterion (eigenvalues > 1) to determine the number of meaningful components. Cronbach's alpha assesses internal consistency of the capacity and ethics subscales, quantifying whether dimensions within each construct reliably measure a coherent latent variable. K-means clustering identifies natural groupings of policies based on their multidimensional profiles, with optimal k determined through silhouette coefficients and stability analysis across bootstrap samples.

4.1.8.7 Hypothesis Testing

Throughout the analyses, we employ both parametric and non-parametric hypothesis tests depending on distributional assumptions. Welch's t-tests compare mean scores between income groups, using the Welch correction to avoid assuming equal variances. Mann-Whitney U tests provide non-parametric alternatives when distributions violate normality assumptions. Chi-square tests assess whether categorical outcomes (such as quadrant membership in the capacity–ethics space) differ by income group. For all tests, we report exact p-values, effect sizes (Cohen's d for mean differences, Cramér's V for categorical associations), and confidence intervals where appropriate, following contemporary standards for transparent statistical reporting.

4.1.9 Reproducibility

All code is available at https://github.com/lsempe77/ai-governance-capacity. The pipeline uses deterministic document IDs (`MD5(url)[:12]`) to ensure reproducibility of the corpus-to-analysis link. API calls to LLM providers used fixed model identifiers and structured JSON output schemas.

4.1.10 Use of Large Language Models

This research employs large language models in two distinct capacities, both of which we disclose here in the interest of methodological transparency.

For data analysis: Large language models (Claude Sonnet 4, GPT-4o, and Gemini Flash 2.0) serve as the core analytical instrument, functioning as automated policy coders that convert unstructured policy documents into structured quantitative scores. This use constitutes the research methodology itself and is documented extensively throughout Section 4.1 and Section 5.1, including validation against human expert ratings. All LLM-generated scores are preserved in the public repository, enabling verification and replication of our analytical pipeline.

For writing assistance: Large language models (primarily GitHub Copilot and Claude) provided assistance with text editing during manuscript preparation. All LLM-generated text was reviewed, revised, and approved by the author, who takes full responsibility for the accuracy and integrity of the final content. LLMs did not generate substantive intellectual contributions, interpret findings, or make analytical decisions — these remained under direct human control throughout the research process.

This dual disclosure reflects our commitment to transparency in an era where LLM use in research is becoming ubiquitous. We distinguish between LLMs as research instruments (where their use is

the methodology being validated) and LLMs as writing assistants (where they augment but do not replace human scholarly judgment).

5 LLM Ensemble Scoring & Validation

5.1 Measuring Governance Quality at Scale

> **Chapter summary.** This chapter presents our LLM-based scoring methodology — a three-model ensemble that independently codes each policy on 10 dimensions. We report inter-rater reliability (ICC = 0.827, Excellent) and discuss model-specific scoring patterns.

5.1.1 Scoring Framework

The transition from collected documents to analyzable data required developing a comprehensive assessment framework that could systematically evaluate implementation readiness across diverse policy types, jurisdictions, and governance traditions. This framework needed to capture both the structural features that enable implementation (capacity dimensions) and the substantive ethical commitments that shape governance outcomes (ethics dimensions). Drawing on decades of implementation science and the emerging AI governance literature, we constructed a 10-dimension assessment framework organized into two complementary domains.

Each of the 2,216 policies was scored on **10 dimensions** using a 0–4 scale, where 0 indicates complete absence of the feature, 1-2 represent minimal to moderate presence, 3 indicates substantial implementation readiness, and 4 reflects comprehensive operationalization with concrete mechanisms. This five-point scale provides sufficient granularity to distinguish meaningful quality differences while maintaining inter-rater reliability — finer scales would introduce excessive noise, while coarser scales would obscure important variation.

5.1.1.1 Capacity Dimensions

Grounded in implementation science (Mazmanian and Sabatier 1983; Lipsky 1980; Grindle 1996; Fukuyama 2013):

Table 5.1: Capacity scoring dimensions

Code	Dimension	What It Measures
C1	Clarity & Specificity	Clear objectives, measurable targets, defined scope
C2	Resources & Budget	Dedicated funding, staffing, infrastructure

Code	Dimension	What It Measures
C3	Authority & Enforcement	Legal mandate, penalties, compliance mechanisms
C4	Accountability & M&E	Reporting, evaluation, oversight bodies
C5	Coherence & Coordination	Cross-agency alignment, international coordination

These five capacity dimensions operationalize the implementation conditions identified by Mazmanian and Sabatier (1983) and extended by subsequent scholars. Clarity corresponds to Mazmanian and Sabatier's emphasis on clear objectives and causal theories; Resources captures Grindle's technical and fiscal capacity requirements; Authority reflects the legal structuring of implementation processes; Accountability operationalizes Lipsky's concern with constraining street-level discretion; and Coherence addresses the coordination challenges documented by Hjern and Hull (1982). Together, they provide a comprehensive assessment of whether policies possess the institutional infrastructure necessary for execution.

5.1.1.2 Ethics Dimensions

Grounded in AI ethics literature (Jobin, Ienca, and Vayena 2019; Floridi et al. 2018; OECD 2019; UNESCO 2021; European Parliament and Council 2024):

Table 5.2: Ethics scoring dimensions

Code	Dimension	What It Measures
E1	Ethical Framework Depth	Grounding in principles, coherent ethical vision
E2	Rights Protection	Privacy, non-discrimination, human oversight, transparency
E3	Governance Mechanisms	Ethics boards, impact assessments, auditing
E4	Operationalisation	Concrete requirements, standards, certification
E5	Inclusion & Participation	Stakeholder processes, marginalised group representation

The ethics dimensions synthesize principles identified across the AI governance literature, particularly the convergence documented by Jobin, Ienca, and Vayena (2019) around transparency, fairness, accountability, and privacy. Framework Depth assesses whether policies ground specific requirements in coherent ethical visions rather than listing buzzwords. Rights Protection operationalizes the human-centric principles emphasized by Floridi et al. (2018) and enshrined in frameworks like UNESCO's AI Recommendation. Governance Mechanisms capture the institutional architecture

for ethics oversight, while Operationalisation distinguishes aspirational statements from concrete requirements with measurable standards. Inclusion reflects the participatory governance emphasis in OECD (2019), recognizing that AI governance legitimacy depends on meaningful stakeholder engagement.

Each dimension uses explicit scoring rubrics (see Section A.1) with anchored examples at each scale point, ensuring that assessments rest on observable textual evidence rather than subjective impressions. Composite scores are computed as unweighted means: *Capacity* = mean(C1–C5), *Ethics* = mean(E1–E5), *Overall* = mean(all 10). This equal weighting reflects our agnostic stance on which dimensions matter most — different governance contexts may prioritize different features, and our framework captures this multidimensionality rather than imposing a single definition of quality.

5.1.2 Three-Model Ensemble

Applying this 10-dimension framework to 2,216 documents requires a scoring approach that balances three competing demands: analytical sophistication (capturing nuanced implementation features), scale (processing millions of words of policy text), and reliability (producing consistent assessments across documents). Traditional human expert coding offers sophistication but becomes prohibitively expensive and time-consuming at this corpus size. Automated keyword-based approaches scale efficiently but lack the interpretive capacity to distinguish substantive implementation details from aspirational rhetoric. Our solution employs frontier large language models as automated policy analysts, leveraging their ability to read and interpret complex documents while maintaining consistency through ensemble design.

To mitigate single-model bias and architectural idiosyncrasies, each policy was independently scored by three frontier LLMs via the OpenRouter API, selected to represent diverse training approaches and institutional origins:

Table 5.3: LLM ensemble composition

Model	Identifier	Role	Entries Scored
Model A	Claude Sonnet 4	Strictest scorer	2,210 (99.7%)
Model B	GPT-4o	Moderate scorer	2,216 (100%)
Model C	Gemini Flash 2.0	Moderate scorer	2,215 (100%)

This ensemble design leverages complementary strengths: Claude Sonnet 4's nuanced policy interpretation and attention to implementation details, GPT-4o's balanced analytical approach and broad domain knowledge, and Gemini Flash 2.0's efficient processing and consistent scoring patterns. By combining models from three different organizations (Anthropic, OpenAI, Google) trained on potentially different corpora using different architectures, we reduce the risk that shared training biases or architectural quirks systematically skew results.

Each model received identical structured prompts containing the full policy text (up to context window limits, typically 8,000+ words) and the complete scoring rubric with anchored examples. The prompts instructed models to read the entire document, assess each dimension independently,

assign a 0-4 score based on observable textual evidence, and provide brief supporting excerpts justifying each score. Models returned structured JSON-formatted outputs with dimension-level scores and evidence, enabling automated aggregation while preserving auditability through the evidence field. The final ensemble score for each dimension is the **median** of the three model scores, following the logic of robust central tendency estimation. The median approach proves superior to the mean in this context because it remains unaffected by single-model outliers and handles the systematic calibration differences we observe across models (detailed below) without requiring explicit recalibration.

The total scoring effort required **6,641 API calls** (2,216 policies × 3 models, minus a handful of failures where models returned malformed JSON or exceeded context windows). The high completion rate — 99.7% of entries successfully scored by all three models — demonstrates the robustness of the pipeline to diverse document formats and lengths.

5.1.3 Inter-Rater Reliability

The validity of this entire analytical enterprise rests on a fundamental question: do the three models agree on policy quality, or do they produce idiosyncratic assessments that reflect model-specific biases rather than genuine document features? If inter-model agreement is low, the ensemble scores become arbitrary — different model combinations would yield different conclusions. If agreement is high, this provides evidence that the scores capture systematic variation in policy quality rather than measurement noise.

We assess agreement across the three LLM "raters" using multiple complementary metrics, following the framework established by Shrout and Fleiss (1979) for inter-rater reliability in observational studies. The intraclass correlation coefficient ICC(2,1) serves as our primary reliability measure, as it appropriately handles the nested structure of our data (three models rating each policy) and quantifies the proportion of total variance attributable to true between-policy differences rather than rater disagreement. We supplement this with pairwise correlations, Fleiss' kappa for categorical agreement, and descriptive measures of score spread to provide a comprehensive reliability portrait.

5.1.3.1 Overall Reliability

Table 5.4: Inter-rater reliability summary

Metric	Value	Interpretation
ICC(2,1) overall	**0.827**	Excellent
ICC(2,1) capacity	0.824	Excellent
ICC(2,1) ethics	0.791	Excellent
Mean pairwise Pearson	0.86	Strong
Mean pairwise Spearman	0.88	Strong
Mean Fleiss'	0.51	Moderate
Mean overall spread	0.40/4	Low disagreement
Scores within 1 point	95.4%	High consistency

Metric	Value	Interpretation

Table 5.4 presents a remarkably consistent picture across multiple metrics. The ICC(2,1) of 0.827 indicates "Excellent" reliability under Cicchetti's (1994) guidelines (>0.75 = Excellent), meaning that approximately 83% of the variance in observed scores reflects true differences between policies rather than rater disagreement. This level of agreement is comparable to or exceeds reliability typically reported in human-coded policy analysis studies, where ICC values of 0.70-0.80 are considered strong evidence of coding quality. The high pairwise correlations (mean r = 0.86, = 0.88) confirm this consistency through a different lens, while the low mean spread (0.40 points on a 4-point scale) and high within-1-point agreement (95.4%) demonstrate that models rarely produce wildly divergent assessments. Even Fleiss' kappa — a more conservative metric that treats the 0-4 scale categorically rather than continuously — achieves moderate agreement (0.51), which for a five-category scale represents substantial consensus.

Crucially, both capacity and ethics subscales achieve excellent reliability independently (ICC = 0.824 and 0.791 respectively), indicating that the strong overall agreement is not driven by a single dominant construct but reflects genuine consensus across both theoretical domains.

5.1.3.2 Dimension-Level ICCs

Table 5.5: Dimension-level ICC values

Dimension	ICC(2,1)	Quality
C1 Clarity	0.720	Good
C2 Resources	0.735	Good
C3 Authority	0.751	Excellent
C4 Accountability	0.753	Excellent
C5 Coherence	0.804	Excellent
E1 Framework	0.751	Excellent
E2 Rights	0.785	Excellent
E3 Governance	0.691	Good
E4 Operationalisation	0.605	Good
E5 Inclusion	0.746	Good

Table 5.5 reveals systematic patterns in dimension-level reliability that illuminate the scoring process. All dimensions achieve at least "Good" reliability (>0.60), with six reaching "Excellent" (>0.75). The highest agreement appears on structural features like Coherence (ICC = 0.804), Authority (0.751), and Rights Protection (0.785) — dimensions where textual evidence is relatively concrete and unambiguous. Lower (though still acceptable) reliability on Operationalisation (0.605) and Governance Mechanisms (0.691) likely reflects the greater interpretive challenge these dimensions pose: distinguishing truly operational requirements from aspirational language requires subtle judgment that even sophisticated models may approach differently. The lowest ICC (E4 Operationalisation, 0.605) still comfortably exceeds conventional acceptability thresholds (>0.40

for exploratory research, >0.60 for established scales), providing confidence that all 10 dimensions contribute meaningful signal rather than noise to the composite scores.

5.1.3.3 Model-Specific Scoring Patterns

The three models exhibit systematic scoring tendencies:

Table 5.6: Model-level mean scores

Model	Capacity Mean	Ethics Mean	Overall Mean
A (Claude)	0.68	0.46	0.57
B (GPT-4o)	0.92	0.71	0.81
C (Gemini)	0.93	0.68	0.81

Table 5.6 exposes a striking and systematic pattern: Model A (Claude Sonnet 4) scores approximately 0.24 points lower on average than Models B and C across both capacity and ethics dimensions. This is not random noise or jurisdiction-specific bias — the pattern holds consistently across all policy types, income groups, and regions, indicating a fundamental calibration difference in how the model interprets the 0-4 scale. Model A appears to require stronger textual evidence to assign higher scores, treating the rubric descriptions more stringently than its counterparts. The gap is particularly pronounced on ethics dimensions (0.46 vs. 0.68-0.71), suggesting that Model A applies more demanding standards for what constitutes operationalized ethical governance versus aspirational principles.

Importantly, this systematic shift does not invalidate Model A's contributions to the ensemble. The high correlation between Model A's scores and those of Models B and C ($r > 0.85$) demonstrates that all three models agree on the *rank ordering* of policies even while disagreeing on absolute levels. The median-based aggregation proves robust to this calibration difference: it preserves the relative rankings while positioning the final scores between the strict and lenient interpretations. An alternative approach using mean scores would require explicit recalibration or standardization; the median avoids this complexity while naturally accounting for systematic shifts.

5.1.3.4 Agreement by Text Quality

Table 5.7: Agreement by text quality

Text Quality	N	Mean Spread	Within 1 pt
Good (500 words)	942	0.57	90.3%
Thin (100–499)	805	0.34	98.9%
Stub (<100)	462	0.13	99.8%

Table 5.7 reveals the expected relationship between document informativeness and scoring consensus. Models achieve near-perfect agreement on stub documents (mean spread 0.13, within-1-point agreement 99.8%), largely because these minimal texts provide insufficient evidence for any dimension to score above zero. The models converge trivially on low scores when documents offer little substance to assess. Agreement remains very high on thin documents (spread 0.34, agreement 98.9%), as these 100-499 word texts typically mention governance features without providing implementation details, again limiting the interpretive range.

The elevated disagreement on good-quality texts (spread 0.57, agreement 90.3%) should not be interpreted as a reliability failure but rather as evidence that models are engaging substantively with document content. Longer, more detailed policies present genuinely ambiguous cases where reasonable analysts might differ: Does a policy with detailed budget projections but unclear enforcement mechanisms score 2 or 3 on Resources? Does sophisticated ethical framework discussion without concrete operationalization merit a 2 or 3 on Framework Depth? These interpretive challenges produce the higher spread we observe. The fact that even for good texts, 90.3% of scores fall within 1 point indicates that disagreement occurs at boundary cases rather than reflecting fundamental divergence in assessment.

5.1.4 Composite Scores

The resulting ensemble produces composite scores with the following distributions:

Table 5.8: Composite score distributions

Component	Mean	SD	Median	IQR
Capacity (C1–C5)	0.83	0.77	0.60	0.00–1.40
Ethics (E1–E5)	0.61	0.62	0.40	0.00–1.00
Overall (all 10)	0.73	0.66	0.50	0.10–1.15

Table 5.8 summarizes the final ensemble scores that serve as the primary data for all subsequent analyses. Three distributional features prove particularly consequential for analytical choices in later chapters.

First, the **strong floor effect** — with 27.6% of policies scoring exactly zero on capacity and 36.3% on ethics — indicates that more than a quarter of documents in the OECD.AI Observatory contain insufficient implementation detail to score above the minimum threshold on our framework. These zeros are not missing data but substantive findings: many AI governance documents consist of brief announcements, aspirational statements, or high-level principles without operational content. This censoring at zero violates the assumptions of standard OLS regression, motivating the Tobit models we employ in Section 7.1 to correct for attenuation bias.

Second, the **right skew** in all three distributions — with medians substantially below means and interquartile ranges concentrated in the lower half of the scale — reveals that most policies cluster at the low end of implementation readiness, while a smaller set of comprehensive policies achieve substantially higher scores. This heterogeneity suggests that focusing solely on mean comparisons

would obscure important distributional differences, motivating the quantile regression approach that examines effects at different points of the score distribution.

Third, the systematic **capacity-ethics gap** — with policies averaging 0.83 on implementation architecture but only 0.61 on ethics operationalization — points to a prioritization pattern: governments more frequently specify institutional structures, budgets, and authorities than operationalize ethical principles through concrete requirements. This gap receives detailed examination in **?@sec-pca-nexus**, where we explore the capacity-ethics nexus and identify distinct governance typologies.

5.1.5 Validation Discussion

The use of large language models as automated policy coders represents a methodological innovation with both promise and peril. Our approach builds on a growing body of evidence demonstrating that frontier language models can perform complex text annotation tasks at or above human-coder quality (Gilardi, Alizadeh, and Kubli 2023; TÃ¶rnberg 2024). Recent validation studies show that LLMs achieve reliability comparable to trained human coders on tasks ranging from sentiment classification to ideological scaling, while processing text orders of magnitude faster and at far lower cost. However, these findings come with important caveats (Pangakis, Wolken, and Fasching 2023): LLM performance varies substantially across task types, prompt formulations, and model versions, and models can exhibit systematic biases learned from training data that may not align with human expert judgment on normatively contentious dimensions.

Three features of our methodological design directly address these validity concerns. The **multi-model ensemble** reduces the risk that findings reflect idiosyncrasies of any single model's training data or architectural choices by combining three independently-developed models from different organizations. If all three models converge on similar assessments despite their different origins, this provides stronger evidence of validity than relying on a single model's output. The **structured output with evidence** requirement — where models must provide supporting textual excerpts justifying each score — enables post-hoc auditing and increases the probability that models ground assessments in observable document features rather than generating plausible-sounding scores without textual basis. The **median aggregation** strategy proves robust both to single-model outliers and to the systematic calibration difference we observe across models, avoiding the need for explicit recalibration while preserving relative rankings.

Important limitations remain that readers should bear in mind when interpreting results. The three models, despite their different origins, may share biases inherited from overlapping training corpora — particularly given that all were likely exposed to prominent AI governance documents like the OECD AI Principles and EU AI Act during training. The scoring rubric itself, while grounded in implementation science theory and AI governance scholarship, necessarily involves subjective judgments about what constitutes "adequate" clarity or "substantial" resource allocation — dimensions on which even expert human coders would reasonably disagree. Our ensemble treats all three models as equally authoritative through median aggregation, but this may not reflect their actual relative validity — it is conceivable that one model's systematic stringency or leniency better aligns with ground truth than the ensemble median, though we lack a gold standard against which to evaluate this.

These methodological uncertainties motivate the extensive robustness checks presented in Section 10.1, where we examine whether core findings hold across alternative specifications, subsamples, and aggregation methods. The consistency of results across these checks provides additional confidence that our conclusions reflect genuine patterns in policy quality rather than artifacts of measurement choices.

6 Capacity Landscape

6.1 The Global Landscape of AI Governance Capacity

> **Chapter summary.** This chapter presents the descriptive landscape of AI governance capacity across 2,216 policies and 70+ jurisdictions. We examine score distributions, income-group comparisons, regional patterns, policy-type variation, and country rankings.

6.1.1 Overall Score Distribution

Before examining differences across countries and policy types, we first establish the baseline landscape: what does AI governance capacity look like globally when all 2,216 policies are considered together? This aggregate view reveals not only central tendencies but also the distributional features that shape subsequent analyses. Understanding the overall landscape proves essential for contextualizing the gaps and clusters we identify in later sections.

The capacity composite score averages **0.83/4.00** (SD = 0.77) across all 2,216 policies — positioning the typical AI governance document substantially below the scale midpoint. Figure 6.1 reveals that this modest mean conceals considerable heterogeneity, with all five dimensions exhibiting pronounced right skew: most policies cluster at or near zero, while a smaller set of comprehensive policies extends into higher score ranges. The dimension-level means range from a low of **0.48** (C4 Accountability) to a high of **1.07** (C5 Coherence):

Table 6.1: 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**

Table 6.1 exposes a striking pattern in the architecture of AI governance globally. **Accountability (C4) is the weakest dimension** across all policies, with a mean of just 0.48 — less than half the strength of Coherence (C5) at 1.07. This gap reflects a systematic prioritization: policies

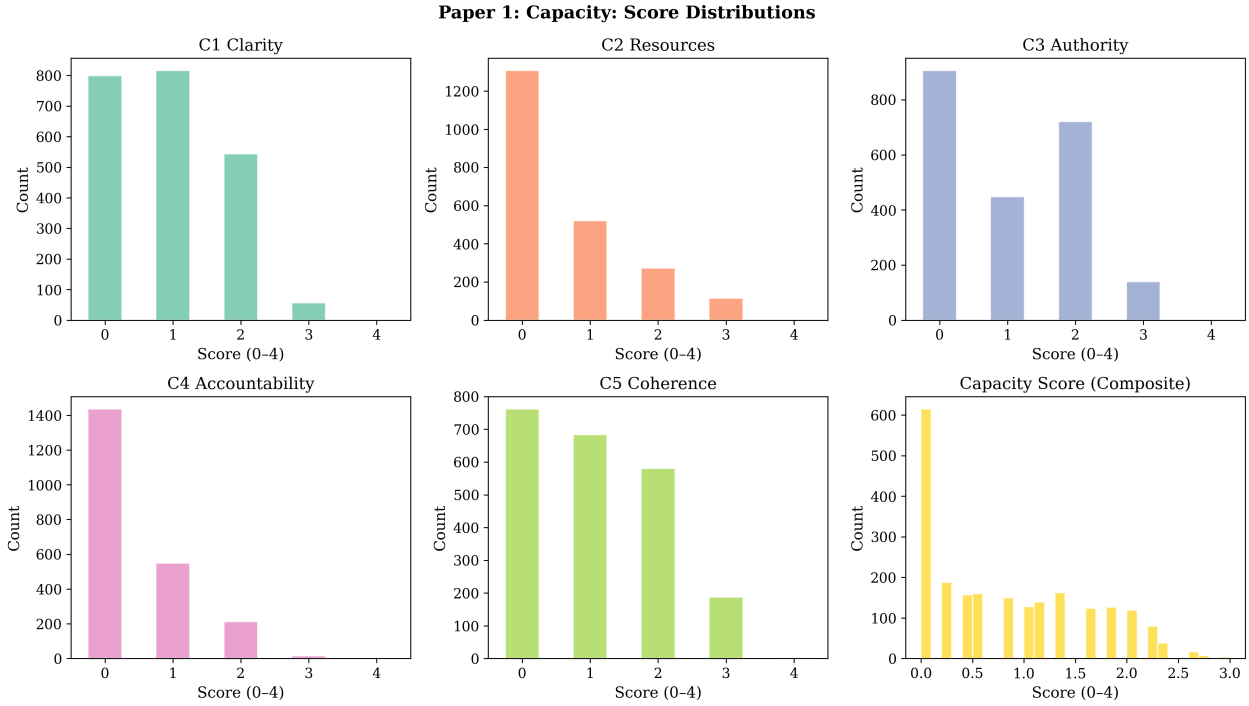

Figure 6.1: Distribution of capacity dimension scores across 2,216 policies. All five dimensions exhibit strong right skew with floor effects at zero.

are more than twice as likely to specify coordination mechanisms or clarify objectives as to establish monitoring and evaluation frameworks. Governments appear more comfortable articulating what they intend to do (Clarity) and how agencies should work together (Coherence) than committing to transparent oversight mechanisms that would enable external assessment of implementation progress.

The substantial standard deviations — ranging from 0.72 to 1.08 — indicate enormous within-dimension variation. A policy scoring 1.0 on Resources might have zero budget allocation or substantial dedicated funding; the heterogeneity spans the full range of implementation readiness. Perhaps most consequentially, **27.6% of all policies score exactly zero** on the capacity composite, indicating no discernible implementation infrastructure in the policy text. More than a quarter of documents in the OECD.AI Observatory function as announcements or aspirational statements rather than operational governance instruments. This floor effect — visible in the median values of 0.00 for Resources and Accountability — motivates the Tobit regression approach we employ in Section 7.1 to correct for censoring bias.

6.1.2 Income-Group Comparisons

The conventional narrative about AI governance assumes a clear North–South divide, with wealthy countries possessing sophisticated regulatory frameworks and developing countries struggling to match this capacity. Our data provide the first systematic test of this assumption across thousands

of policies. The results prove more nuanced than this binary narrative suggests: while a gap exists in the raw data, its magnitude and robustness merit careful examination.

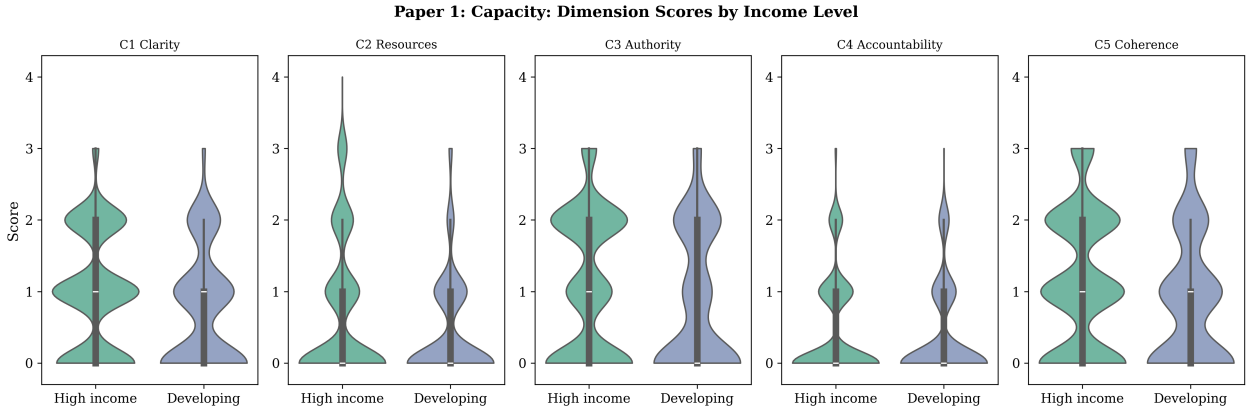

Figure 6.2: Violin plots comparing capacity score distributions between high-income and developing countries. The overlap between distributions is substantial.

Figure 6.2 reveals both the gap and its limitations visually. High-income countries score significantly higher on capacity (mean 0.87, SD 0.77) than developing countries (mean 0.65, SD 0.72), a difference that proves statistically significant by conventional standards:

Table 6.2: 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

Table 6.2 presents the conventional statistical evidence for an income-based capacity gap. The Welch's t-test yields t = 5.47, p < .001, providing strong statistical evidence against the null hypothesis of equal means. However, statistical significance at large sample sizes (N = 2,097) does not automatically imply substantive importance. The Cohen's d effect size of **0.30** falls into the "small" range by conventional standards, indicating that the distributions overlap considerably. Indeed, the violin plots show that many developing-country policies score above the high-income median, while many high-income policies cluster near zero.

Crucially — and foreshadowing findings detailed in Section 10.1 — this gap **vanishes entirely** when analyses are restricted to well-documented policies with good text quality (d = 0.04, n.s.). This sensitivity to text extraction quality raises important interpretive questions: Does the observed gap reflect genuine capacity differences, or does it emerge from systematic differences in how countries document their policies? Do developing countries produce shorter policy documents because

they have less capacity to document, or because their policy dissemination practices differ? This measurement challenge proves central to understanding what the capacity gap means.

6.1.2.1 Dimension-Level Gaps

The aggregate capacity gap masks considerable heterogeneity across the five implementation dimensions. If the North–South divide reflected fundamental differences in state capacity, we would expect relatively uniform gaps across all dimensions. Instead, the pattern proves more differentiated, suggesting that specific capacity constraints — rather than generalized institutional weakness — drive observed differences.

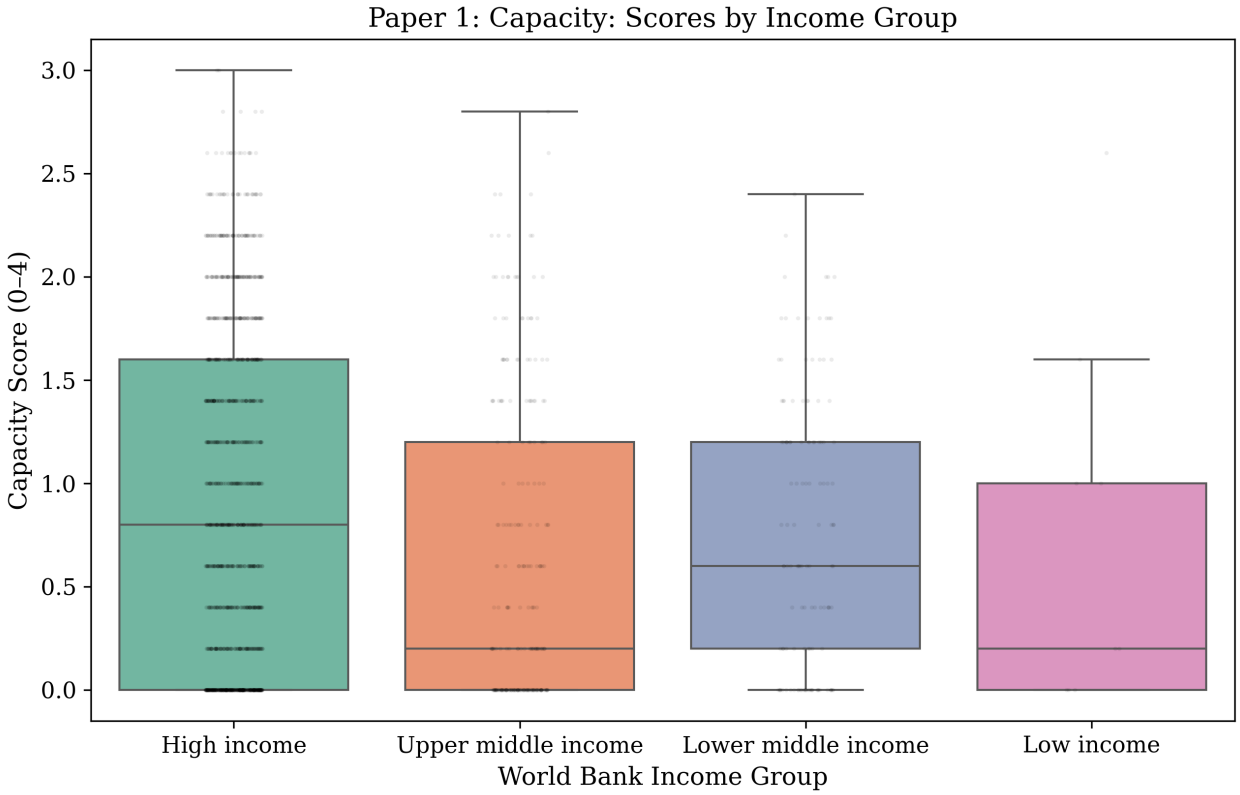

Figure 6.3: Boxplots of capacity scores by income group across all five dimensions.

Figure 6.3 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:

Table 6.3: 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**	$< .001$

Dimension	HI Mean	Dev Mean	Diff	d	p
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	$< .001$

Table 6.3 confirms hypothesis H3: the capacity gap proves largest in **Resources (C2)** (d = 0.32) and smallest in **Accountability (C4)** (d = 0.15). This pattern makes intuitive sense when considered through the lens of implementation science. Specifying budgets, staffing plans, and technical infrastructure requires fiscal resources that correlate with national wealth — wealthy countries can commit larger absolute budgets and possess deeper pools of technical expertise to deploy. By contrast, designing monitoring and evaluation frameworks represents primarily a policy design choice rather than a resource constraint — developing countries can establish reporting requirements, evaluation mandates, and oversight bodies with minimal fiscal commitment.

The modest gap on Accountability carries an ironic implication: developing countries could relatively easily narrow the capacity divide by strengthening their weakest dimension, yet both income groups underperform on accountability mechanisms. The universally low C4 scores (0.48 for high-income, 0.37 for developing) suggest that the reluctance to establish transparent oversight mechanisms transcends wealth differences. Accountability frameworks create political risks by enabling external assessment of implementation failures — a concern that affects governments regardless of income level.

6.1.3 Regional Patterns

Income-group comparisons, while revealing aggregate patterns, risk obscuring important geographic heterogeneity within income categories. The developing-country category encompasses Latin American middle-income countries with sophisticated regulatory traditions, South Asian nations with large technology sectors but limited governance infrastructure, and Sub-Saharan African countries with nascent AI policy ecosystems. Similarly, the high-income group conflates North American and European regulatory leaders with smaller high-income countries that have produced minimal AI governance activity. Regional analysis provides finer-grained insight into governance patterns that income alone cannot capture.

Figure 6.4 reveals that regional variation is substantial but not straightforwardly reducible to income differences. Several patterns challenge simplistic assumptions about governance capacity. **North America (NAM)** leads across all five dimensions, driven primarily by the United States and Canada's extensive policy portfolios and consistently high-scoring individual documents. The North American strength appears most pronounced on Resources (C2) and Accountability (C4) — dimensions where fiscal capacity and regulatory tradition matter most.

Europe & Central Asia (ECA) shows the broadest dimensional coverage, with particular strength in Coherence (C5) reflecting the European Union's multilevel coordination mechanisms. The EU's extensive cross-border governance infrastructure — including the AI Act, Digital Services Act, and GDPR — establishes coordination frameworks that national policies reference and build

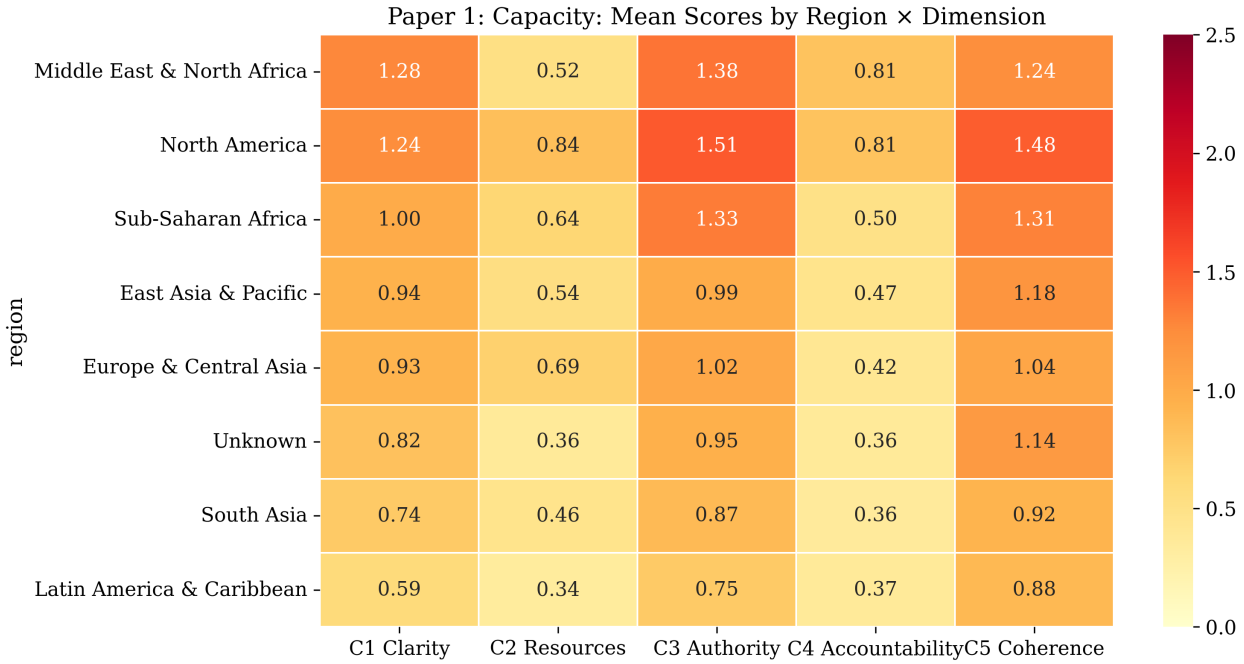

Figure 6.4: Heatmap of mean capacity scores by region and dimension. North America and Europe & Central Asia lead; Sub-Saharan Africa and South Asia trail.

upon. This regional coordination advantage distinguishes Europe from other high-income regions where governance remains more fragmented across national boundaries.

Perhaps most striking are the developing regions that exceed income-predicted performance. **Latin America & Caribbean (LAC)** scores above its income-group average across multiple dimensions, with particular strength in Authority (C3). Countries like Brazil, Colombia, and Argentina have adopted binding AI legislation with clear enforcement mechanisms, demonstrating that regulatory sophistication does not require first-world wealth. **Sub-Saharan Africa (SSA)**, while scoring lowest overall, shows surprising strength in Authority — several African countries (Kenya, Rwanda, South Africa) have enacted AI-specific legislation with legal mandates and compliance mechanisms that exceed what many wealthy countries have adopted.

These regional patterns suggest that governance capacity reflects institutional and political factors beyond simple wealth accumulation. Regulatory traditions, regional coordination frameworks, and policy diffusion networks shape capacity in ways that income alone cannot explain. We return to these themes in Section 9.1, where we examine how policies spread across countries and regions.

6.1.4 Policy-Type Variation

Not all policy documents serve the same function or possess the same implementation obligations. National AI strategies articulate long-term visions and coordinate across sectors but may lack enforcement teeth. Binding regulations establish legal requirements with compliance mechanisms but may sacrifice flexibility. Ethics guidelines provide normative frameworks without operational

detail. These functional differences suggest that capacity scores should vary systematically by policy type — not because some jurisdictions lack capacity but because different document types serve different governance purposes.

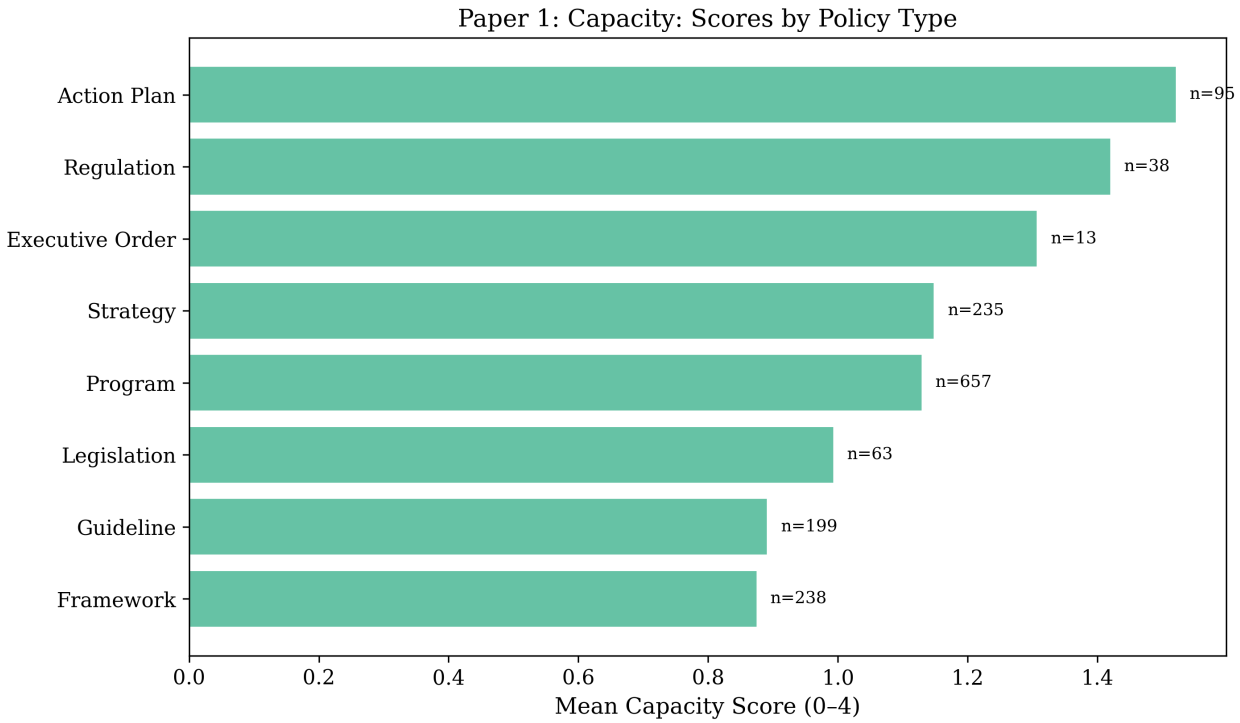

Figure 6.5: Capacity scores by policy type. Binding regulations score highest; guidelines and principles score lowest.

Figure 6.5 confirms this intuition clearly. **Binding regulation** — including laws, executive orders, and regulatory frameworks — scores highest across all dimensions, with particularly pronounced advantages on Authority (C3) and Accountability (C4). This makes structural sense: legal instruments must specify enforcement mechanisms, define responsible agencies, and establish monitoring procedures to be judicially enforceable. The high Authority scores reflect the legal mandates that binding regulations inherently possess, while elevated Accountability scores indicate that laws more frequently establish reporting requirements and evaluation frameworks than softer policy instruments.

National strategies occupy the middle ground, scoring moderately across dimensions with relative strength in Clarity (C1) and Coherence (C5). Strategic documents excel at articulating clear objectives and coordinating across government agencies — their primary intended function — while naturally scoring lower on Resources and Authority, which strategies often leave to subsequent implementing legislation. This pattern validates our measurement approach: scores reflect document content appropriate to policy type rather than applying a one-size-fits-all standard.

Guidelines and principles score lowest, reflecting their aspirational rather than operational character. Ethics guidelines typically enumerate desirable AI properties (transparency, fairness, accountability) without specifying who must implement these principles, how compliance will be monitored,

or what resources will be allocated. These documents serve important functions — establishing normative frameworks, coordinating international principles, guiding voluntary adoption — but they deliberately avoid the operational specificity that would generate high capacity scores. The low scores on guidelines thus reflect accurate measurement of their non-binding character rather than a failure of governance.

This policy-type variation carries important implications for cross-national comparison. Jurisdictions that rely primarily on voluntary guidelines will systematically score lower than those adopting binding legislation, even if the voluntary approach proves effective through industry self-regulation. Our subsequent analyses control for policy type to ensure that apparent capacity differences reflect genuine institutional variation rather than strategic choices about governance instruments.

6.1.5 Country Rankings

Aggregate statistics and distributional analyses provide essential context, but policymakers and researchers often want direct answers to a simpler question: which countries demonstrate the strongest AI governance capacity, and what distinguishes leaders from laggards? Country-level rankings compress our multidimensional framework into a single evaluative scale, inevitably losing nuance but gaining interpretive clarity. These rankings reflect both the number of policies each jurisdiction has produced and the average quality of those policies — a country with five excellent policies will rank higher than one with ten mediocre ones.

Figure 6.6 shows the evolution of capacity scores over time, revealing that aggregate capacity has remained relatively stable since 2019 despite the proliferation of policies. This stability suggests that policy quantity has not translated into quality improvement — the rapid expansion of AI governance activity has produced many low-scoring documents alongside continued high-quality policy development by leading jurisdictions.

The top-scoring jurisdictions combine large policy portfolios with consistently high-quality individual policies, demonstrating sustained commitment to implementation-ready governance rather than one-off symbolic gestures:

Table 6.4: 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

Table 6.4 reveals an unsurprising but important pattern: the top four positions are occupied by high-income jurisdictions with extensive AI policy ecosystems and sophisticated regulatory traditions. The **European Union** leads with a mean capacity score of 1.42, reflecting its comprehensive regulatory framework anchored by the AI Act and supported by extensive complementary policies. **Canada** (1.38) and the **United Kingdom** (1.32) demonstrate that smaller policy portfolios can

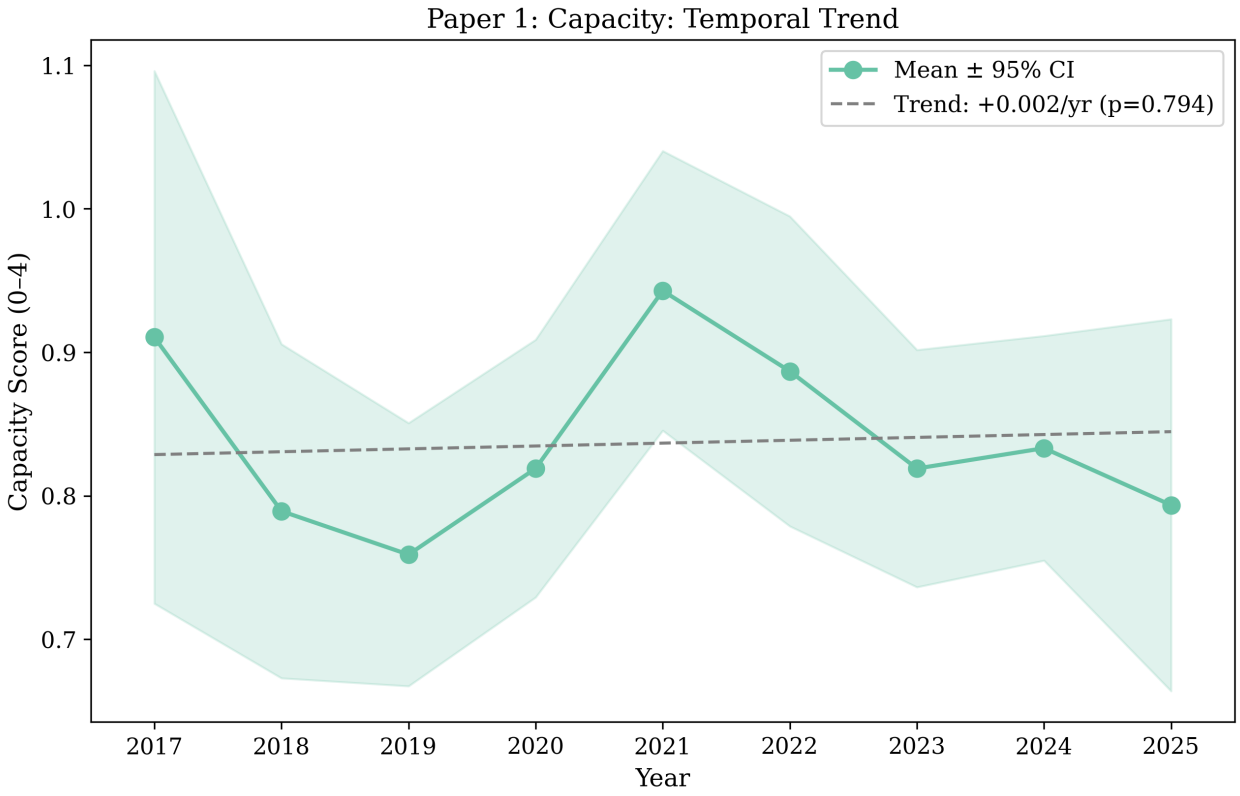

Figure 6.6: Temporal trends in capacity scores across the 2017–2025 period.

achieve high average quality through careful policy design. The **United States** (1.28), despite the largest policy portfolio (84 documents), ranks fourth due to considerable variation across federal agencies and states — some US policies achieve top-tier scores while others function as brief announcements.

The fifth position, however, disrupts this high-income dominance. **Colombia** (1.21), an upper-middle-income country, outperforms numerous wealthy nations through a focused portfolio of eight well-designed policies. Colombia's achievement exemplifies a pattern explored extensively in Section 9.1: several developing countries punch above their GDP-predicted weight through strategic policy choices rather than resource abundance. Notable developing-country performers beyond the top five include **Brazil** (consistently in the top 10) and **Kenya** (ranking above many European countries despite substantially lower GDP per capita).

These outliers prove analytically valuable precisely because they challenge the assumed tight coupling between wealth and governance capacity. If Colombia, Brazil, and Kenya can achieve strong implementation readiness despite resource constraints, this suggests that institutional design choices — clarity of objectives, establishment of coordination mechanisms, specification of authorities — matter more than fiscal capacity alone. We return to these efficiency frontier countries in subsequent chapters to understand what enables their outperformance.

6.1.6 Correlation Structure

The preceding analyses treat the five capacity dimensions as separate constructs, examining how Clarity differs from Resources or how Accountability gaps vary by income. But are these truly distinct dimensions, or do they simply measure the same underlying governance quality factor with slightly different labels? The correlation structure among dimensions reveals whether our framework captures meaningful multidimensionality or redundantly measures a single latent construct.

Figure 6.7 shows that the five capacity dimensions are indeed positively correlated, with pairwise correlations ranging from $r = 0.45$ to $r = 0.75$. This pattern indicates substantial shared variance — policies that score high on one dimension tend to score high on others, suggesting a common underlying governance quality factor. The strongest correlations link Authority (C3) with Coherence (C5) ($r = 0.75$), and Clarity (C1) with Authority ($r = 0.70$), reflecting natural implementation logic: policies with clear legal mandates more readily establish coordination mechanisms, and those with specific objectives more effectively define enforcement authorities.

However, the correlations remain well below 1.0, indicating that the dimensions maintain sufficient discriminant validity to justify separate measurement. A policy can score high on Clarity (well-defined objectives) while scoring low on Resources (no budget allocation), or achieve strong Coherence (cross-agency coordination) despite weak Accountability (no monitoring mechanisms). The moderate correlation structure suggests that our framework successfully captures distinct facets of implementation capacity rather than redundantly measuring a single dimension.

This structure receives formal confirmation through the principal component analysis presented in **?@sec-pca-nexus**, which demonstrates that while capacity dimensions share approximately 66% of their variance through a general governance factor, they also exhibit sufficient independence to warrant separate analysis. The multidimensional framework thus provides richer diagnostic

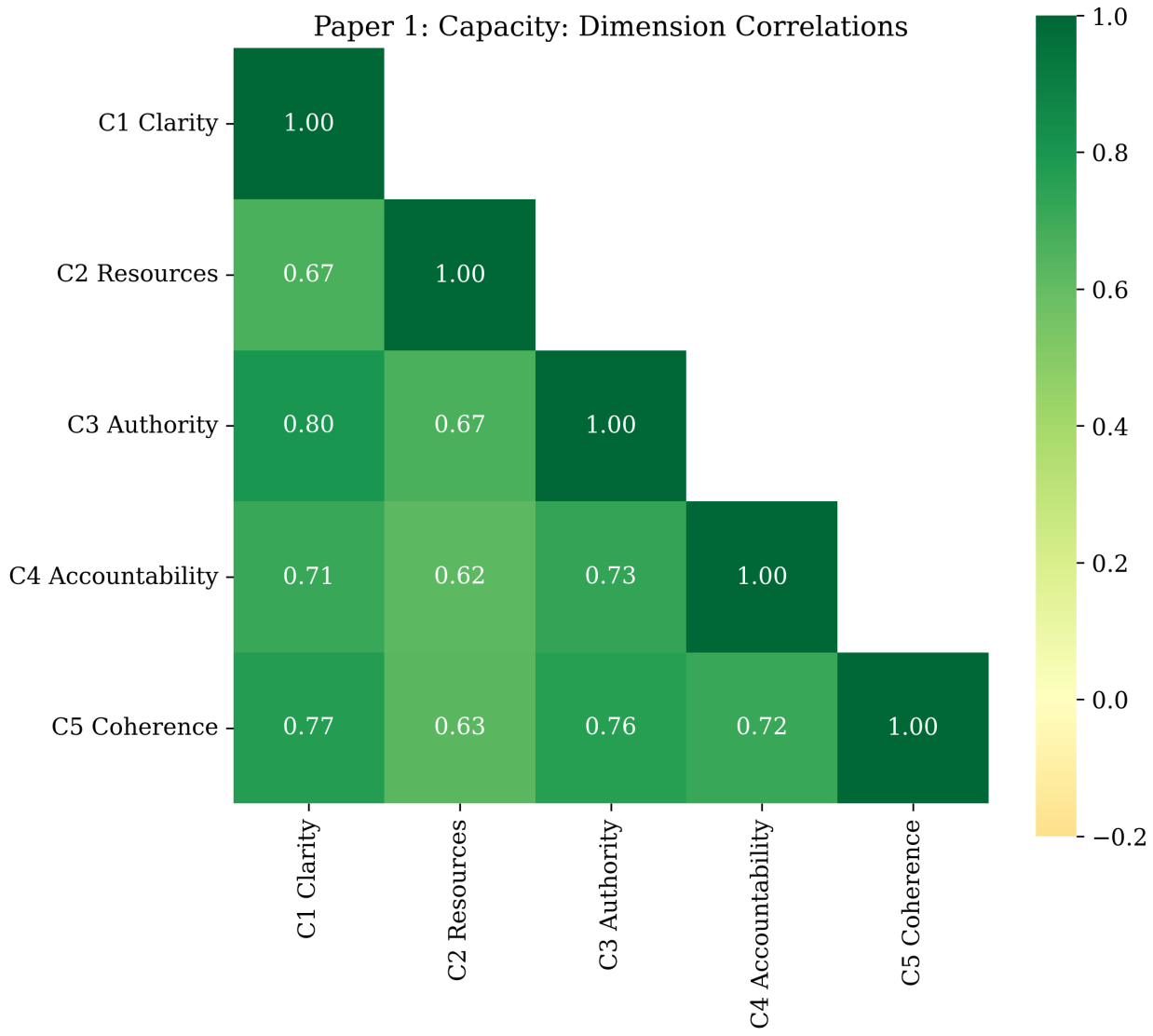

Figure 6.7: Correlation matrix across the five capacity dimensions. All dimensions are positively correlated, with the strongest link between C3 (Authority) and C5 (Coherence).

information than a single overall quality score would permit, enabling identification of specific capacity strengths and weaknesses that aggregate scores would obscure.

7 Capacity Determinants

7.1 What Explains Governance Capacity?

Chapter summary. This chapter examines the determinants of AI governance capacity through four complementary regression approaches: OLS, multilevel models, quantile regression, and Tobit models. The central finding is that GDP per capita explains remarkably little variation — text quality dominates.

7.1.1 OLS Regression

Having established the descriptive landscape of AI governance capacity — its distribution, income-group gaps, regional patterns, and country rankings — we now turn to the analytical question of *explanation*: what factors determine why some policies demonstrate high implementation readiness while others remain aspirational? The conventional expectation holds that national wealth drives governance capacity, with wealthy countries possessing the fiscal resources, technical expertise, and institutional infrastructure necessary for sophisticated AI regulation. Testing this hypothesis requires moving beyond descriptive comparison to multivariate regression that can isolate the effect of GDP while controlling for other predictors of capacity.

Figure 7.1 provides the first visual evidence, plotting capacity scores against log GDP per capita. While the relationship appears positive — the regression line slopes upward — the scatter around this line proves substantial. Many developing-country policies score above the line while many high-income policies score below it, suggesting that wealth alone provides limited predictive power. The formal test employs a standard OLS model predicting the capacity composite:

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

Table 7.1: 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	$< .001$

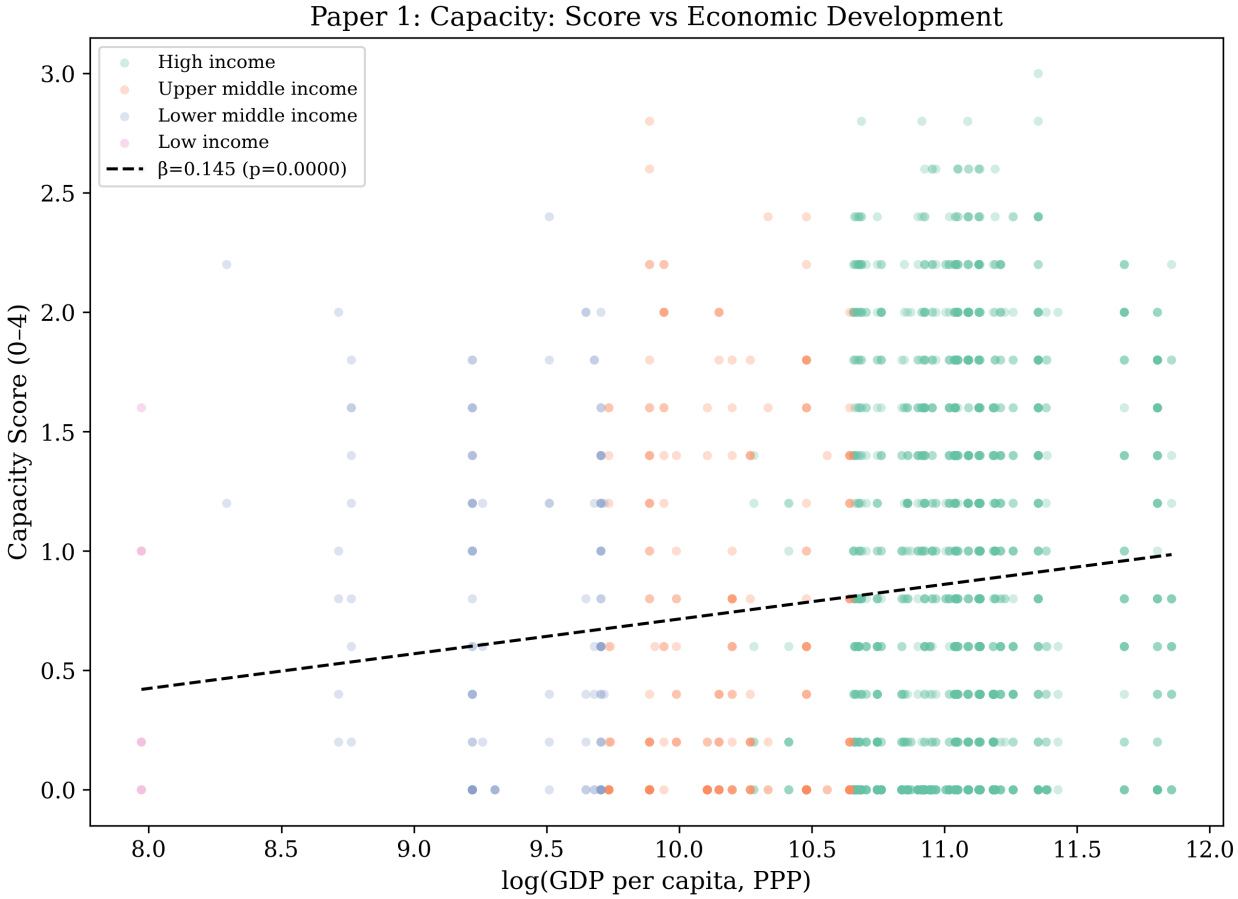

Figure 7.1: Scatter plot of capacity scores against log GDP per capita, with regression line. The relationship is positive but weak.

Table 7.1 reveals three critical findings that fundamentally shape interpretation of governance capacity gaps. First, **text quality dominates** all other predictors by an enormous margin. The coefficient on `is_good_text` (= 1.004, t = 37.64) dwarfs every substantive predictor, indicating that policies with at least 500 words of extracted text score a full point higher on the 0-4 scale than those with minimal documentation. This massive effect represents primarily a measurement artifact rather than a genuine governance difference: longer documents provide more textual evidence for LLM coders to assess, mechanically generating higher scores. The fact that text quality alone produces a t-statistic of 37.64 — far exceeding any conventional threshold — demonstrates its overwhelming influence on observed scores.

Second, **GDP matters, but only modestly**. The log GDP per capita coefficient (= 0.086, p < .001) achieves statistical significance but substantive importance proves limited. A one-unit increase in log GDP — roughly equivalent to tripling a country's wealth from $10,000 to $30,000 per capita — raises capacity scores by only 0.086 points on a 4-point scale. This means moving from a typical developing-country income level to a typical high-income level would increase capacity scores by less than one-tenth of a point, a difference imperceptible against the backdrop of within-group variation we documented in Section 8.1. The weak GDP effect challenges the conventional assumption that wealth automatically translates into governance capacity.

Third, **binding regulation adds capacity** through its inherent operational character. Laws and executive orders score 0.19 points higher than soft-law instruments (= 0.190, p = .006), reflecting the structural reality that legally enforceable policies must specify authorities, enforcement mechanisms, and monitoring procedures to be judicially defensible. This effect validates our measurement approach: scores appropriately reflect document content matched to policy type rather than applying uniform standards regardless of function.

The model's overall fit provides the most consequential finding: $R^2 = 0.436$, indicating that 44% of variance in capacity scores can be explained by these predictors. However, this seemingly respectable fit conceals a troubling reality. When text quality is removed from the model, R^2 collapses to 0.012 — meaning that GDP, policy year, and document type combined explain only 1.2% of variation in capacity scores. Nearly all the model's explanatory power derives from the measurement artifact of text length rather than substantive governance differences. This sensitivity to text quality motivates the robustness checks in Section 10.1, where we restrict analyses to well-documented policies to determine whether the income gap persists when measurement quality is held constant.

7.1.2 Multilevel Models

The OLS specification rests on a critical assumption: that observations are independent, with each policy's score unaffected by others in the dataset. This assumption fails in our context because policies nest within countries. Multiple policies from the same jurisdiction share institutional features, regulatory traditions, and political contexts that the OLS model ignores. This clustering creates dependency: policies from Canada may consistently score higher or lower than the overall mean due to Canadian-specific factors, and knowing one Canadian policy's score provides information about others. Ignoring this nested structure produces two problems: standard errors become artificially small (inflating statistical significance), and country-level effects get incorrectly attributed to policy-level predictors.

The solution employs multilevel modeling with random intercepts for countries, allowing each jurisdiction to have its own baseline capacity level around which individual policies vary. This approach explicitly partitions variance into between-country and within-country components, revealing how much of the apparent GDP effect reflects genuine cross-national differences versus policy-level variation.

We estimate a random-intercept multilevel model:

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

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

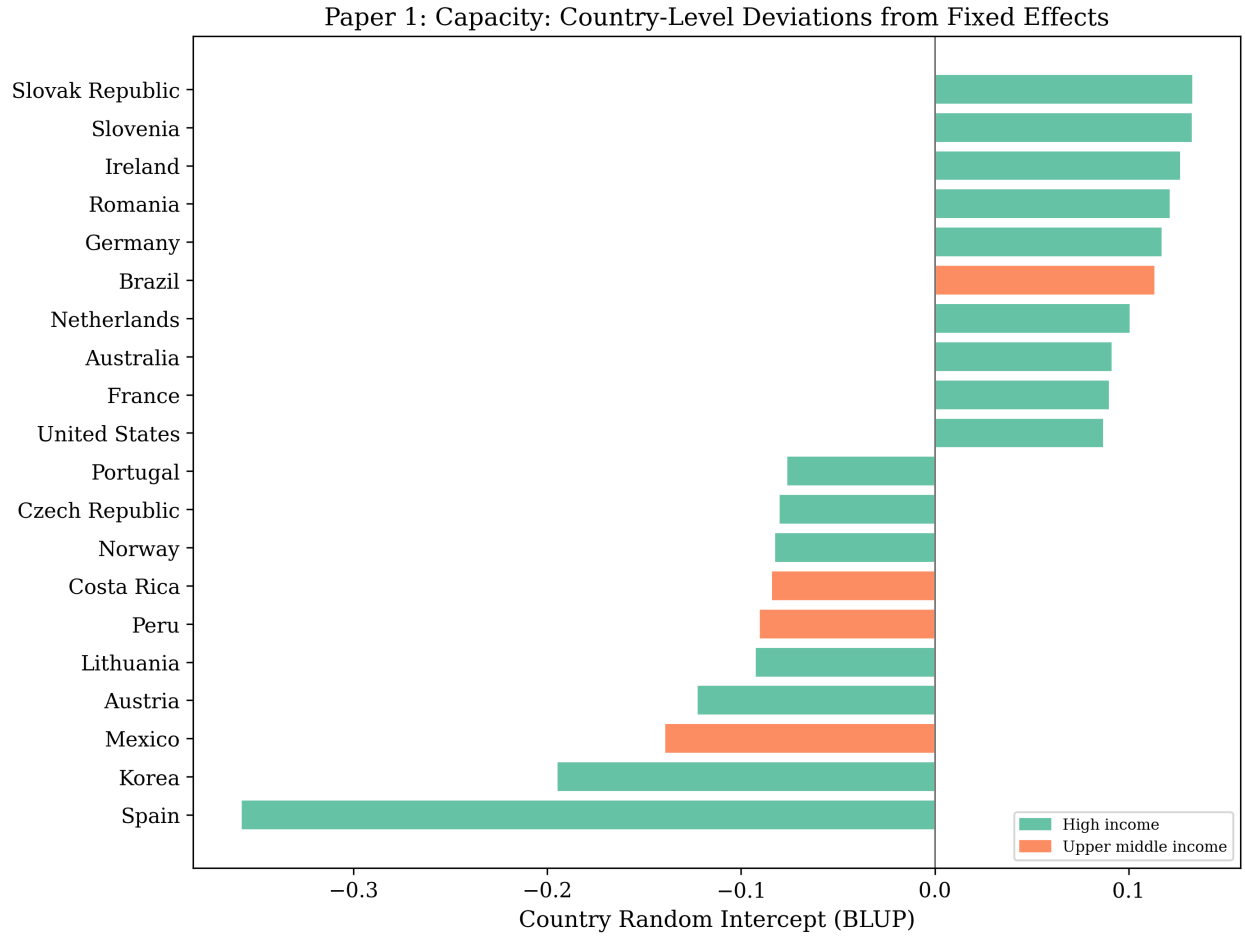

Figure 7.2: Country random effects from the multilevel capacity model. Most countries cluster near zero, with a few notable over- and under-performers.

Table 7.2: 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	—	$p = .007$

Table 7.2 exposes the consequences of ignoring nested structure. The **country-level ICC = 0.091** delivers a striking finding: only 9.1% of total variance occurs between countries, while 90.9% occurs between policies within countries. This decomposition fundamentally challenges wealth-based explanations of governance capacity. If national characteristics — GDP, institutions, political systems — primarily determined capacity, we would expect substantial between-country variance as countries cluster around characteristic levels. Instead, the overwhelming dominance of within-country variance (91%) indicates that policy-specific features matter far more than country-level attributes. A Canadian policy is more different from the average Canadian policy than the average Canadian policy is from the average Brazilian policy.

The multilevel specification reveals that **OLS inflates the GDP effect** by approximately 25%. The mixed-model GDP coefficient (= 0.066) proves substantially smaller than the OLS estimate (= 0.088), because OLS double-counts country-level variation. When a country like Denmark has many high-scoring policies, OLS attributes this pattern entirely to Denmark's high GDP, ignoring the possibility that multiple Danish policies score high for Denmark-specific reasons unrelated to wealth. The multilevel model correctly partitions these sources, revealing that the GDP effect is even weaker than OLS suggests.

The likelihood ratio test confirms that **the multilevel model provides the correct specification** (LR test p = .007), formally rejecting the OLS assumption of independence. However, the practical difference remains modest — the GDP coefficient declines from 0.088 to 0.066, neither value suggesting strong wealth effects. The more consequential finding emerges from the variance decomposition: governance capacity is overwhelmingly a policy-level rather than country-level phenomenon.

7.1.3 Quantile Regression

Both OLS and multilevel models estimate mean effects: how does GDP affect the *average* policy's capacity score? This focus on central tendency obscures potential heterogeneity across the score distribution. GDP might matter enormously for moving policies from zero to minimal capacity (getting basic institutional architecture in place) while proving irrelevant for distinguishing good policies from excellent ones. Alternatively, wealth might primarily help countries produce sophisticated policies at the high end while having little effect on preventing low-quality outputs. Standard regression cannot detect such distributional heterogeneity because it collapses all policies into a single average effect.

Quantile regression (Koenker and Bassett 1978) addresses this limitation by estimating separate effects at different points of the distribution — the 25th percentile (low-scoring policies), the median (typical policies), and the 75th percentile (high-scoring policies). This approach reveals whether GDP effects vary across the capacity spectrum, providing diagnostic insight into where wealth matters most for governance quality.

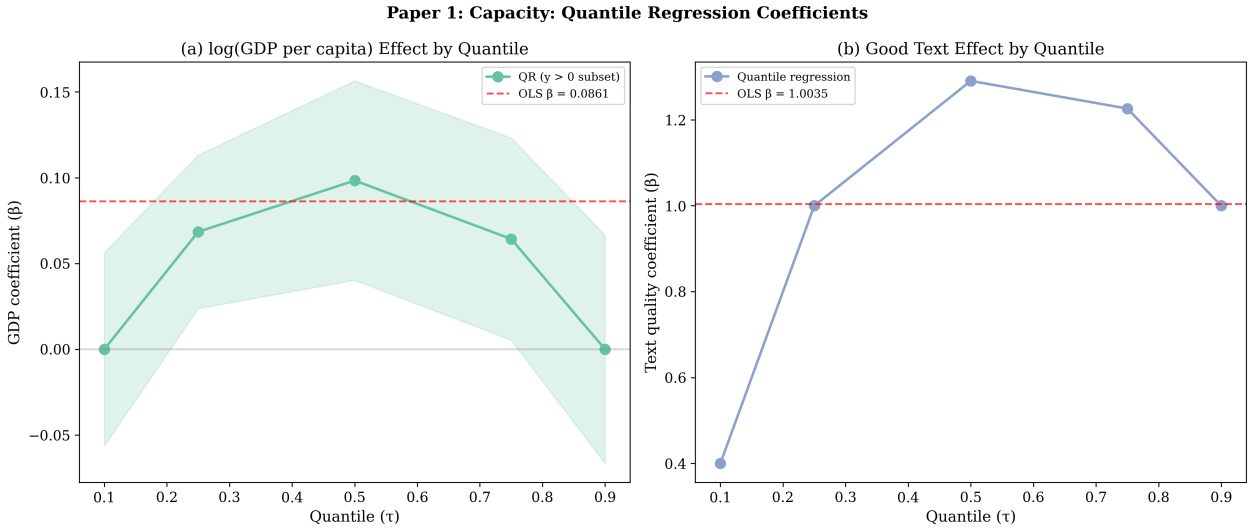

Figure 7.3: Quantile regression coefficients for GDP across the capacity distribution. GDP matters at the median but not at the extremes — an inverted-U pattern.

Figure 7.3 visualizes the GDP effect across quantiles, revealing a striking **inverted-U pattern** where GDP matters most at the median and least at both tails. The formal estimates quantify this heterogeneity:

Table 7.3: 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	$< .001$

Table 7.3 confirms the inverted-U pattern statistically. The GDP effect proves **strongest at the median** (= 0.50, = 0.098, p $< .001$), where typical policies cluster, and **weaker at both extremes**. At the 25th percentile (= 0.068) and 75th percentile (= 0.064), GDP effects are 30-35% smaller than at the median, though they remain statistically significant.

This pattern admits a straightforward interpretation grounded in the mechanisms linking wealth to capacity. At the lowest quantiles, many policies score zero regardless of national GDP because they consist of brief announcements or aspirational statements without operational content. No amount of national wealth can transform a press release into an implementation plan — these minimal

documents systematically score near zero independent of country resources. The floor effect thus compresses the lower tail, limiting GDP's influence.

At the highest quantiles, top-performing policies achieve comprehensive scores through factors beyond fiscal resources: policy learning from international models, political commitment to robust governance, technical sophistication in policy design, and effective stakeholder consultation. Countries can produce excellent policies through these pathways regardless of wealth — Brazil's AI governance framework demonstrates that sophisticated design does not require first-world GDP. The ceiling effect thus loosens the wealth-capacity link at the high end, as governance sophistication becomes primarily a choice variable rather than a resource constraint.

GDP matters most in the middle range, where policies transition from minimal to moderate implementation readiness. Wealthy countries more reliably avoid zero-scoring announcements and more frequently reach moderate capacity levels with some operational detail, resource specification, and enforcement structure. But this advantage proves temporary: achieving truly high capacity requires factors orthogonal to wealth, explaining why the GDP effect declines again at the 75th percentile. The inverted-U thus reveals that wealth primarily helps countries avoid governance failures rather than ensuring governance excellence.

7.1.4 Tobit Regression

The quantile regression revealed that GDP effects weaken at the lower tail where floor effects concentrate. This finding points to a broader statistical challenge: the strong floor effect documented in Section 6.1, where 27.6% of policies score exactly zero, violates OLS assumptions. Standard regression treats zero as just another value along a continuous scale, but scores cannot fall below zero — the distribution is censored at the lower bound. When censoring is severe, OLS produces attenuated coefficient estimates that understate true effects, because it incorrectly treats observed zeros as the latent capacity these policies would have exhibited absent the floor constraint.

Tobit regression (Tobin 1958) explicitly models this left-censoring, distinguishing latent capacity (which can be negative, indicating policies that would score below zero if the scale permitted) from observed capacity (truncated at zero). The model estimates what capacity scores would be absent the floor constraint, then applies the censoring rule to generate predicted observations. This approach yields unbiased coefficient estimates when censoring is present.

The Tobit model with left-censoring at zero takes the form:

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

Table 7.4: 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×

Variable	OLS β	Tobit β	Ratio
σ	—	0.742	—

Table 7.4 and Figure 7.5 demonstrate that OLS indeed attenuates the GDP effect in the presence of floor censoring. The Tobit GDP coefficient (= 0.121) proves **41% larger** than the OLS estimate (= 0.086), confirming that treating zeros as typical observations understates the true wealth-capacity relationship. The estimated error standard deviation = 0.742 quantifies residual variation after accounting for censoring, while the probability of being uncensored at the mean predictor values (82.7%) indicates that censoring affects more than a quarter of observations — substantial enough to bias OLS but not so extreme that most observations cluster at zero.

Crucially, however, even the corrected Tobit coefficient of 0.121 remains substantively modest. Tripling GDP per capita increases capacity by only 0.12 points on a 4-point scale — enough to move from "minimal" to "low-moderate" implementation readiness but insufficient to explain the full range of variation we observe. The correction for censoring matters methodologically, confirming that OLS underestimates effects, but it does not fundamentally alter the substantive conclusion that GDP provides limited explanatory power for governance capacity.

The text quality coefficient experiences similar inflation (from 1.004 in OLS to 1.193 in Tobit), reinforcing that measurement artifacts dominate substantive predictors even after correcting for floor effects. The binding regulation coefficient, by contrast, shrinks slightly (from 0.190 to 0.174), suggesting that document type effects prove somewhat sensitive to specification choices — though the direction and significance remain stable across models.

7.1.5 Synthesis

Across four complementary estimation strategies — each addressing different threats to valid inference — three consistent findings crystallize that fundamentally shape how we interpret AI governance capacity.

First, **text quality dominates all substantive predictors** across every specification. Whether we employ OLS, multilevel models, quantile regression, or Tobit estimation, the coefficient on document length dwarfs GDP, policy type, and temporal trends. This measurement artifact accounts for more variance than all theoretically-motivated predictors combined, indicating that much of what appears as governance quality actually reflects documentation practices. The robustness checks in Section 10.1 examine whether substantive findings persist when we restrict analyses to well-documented policies, effectively holding measurement quality constant.

Second, **GDP matters, but only modestly**, even after every methodological correction. The multilevel model accounts for clustered observations, reducing the GDP coefficient by 25%. The Tobit model corrects for floor censoring, inflating the coefficient by 41%. Yet across all specifications, GDP per capita explains only 3-12% of variance in governance scores, with point estimates suggesting that tripling national wealth increases capacity by 0.07-0.12 points on a 4-point scale. This weak relationship fundamentally challenges the assumption that wealth determines governance capacity. GDP is not destiny — institutional choices, policy design sophistication, and political commitment prove far more consequential than national income.

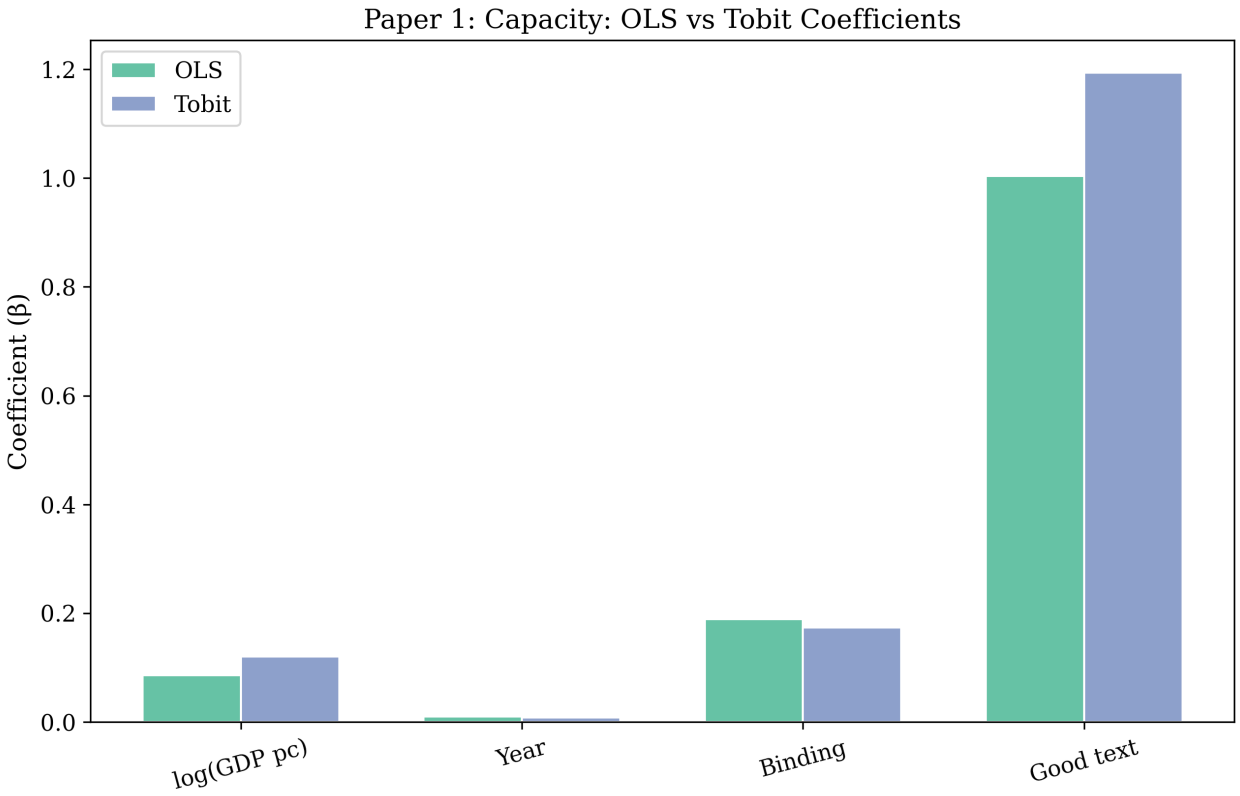

Figure 7.4: Comparison of OLS and Tobit coefficients for the capacity model. Tobit coefficients are systematically larger, reflecting correction for censoring.

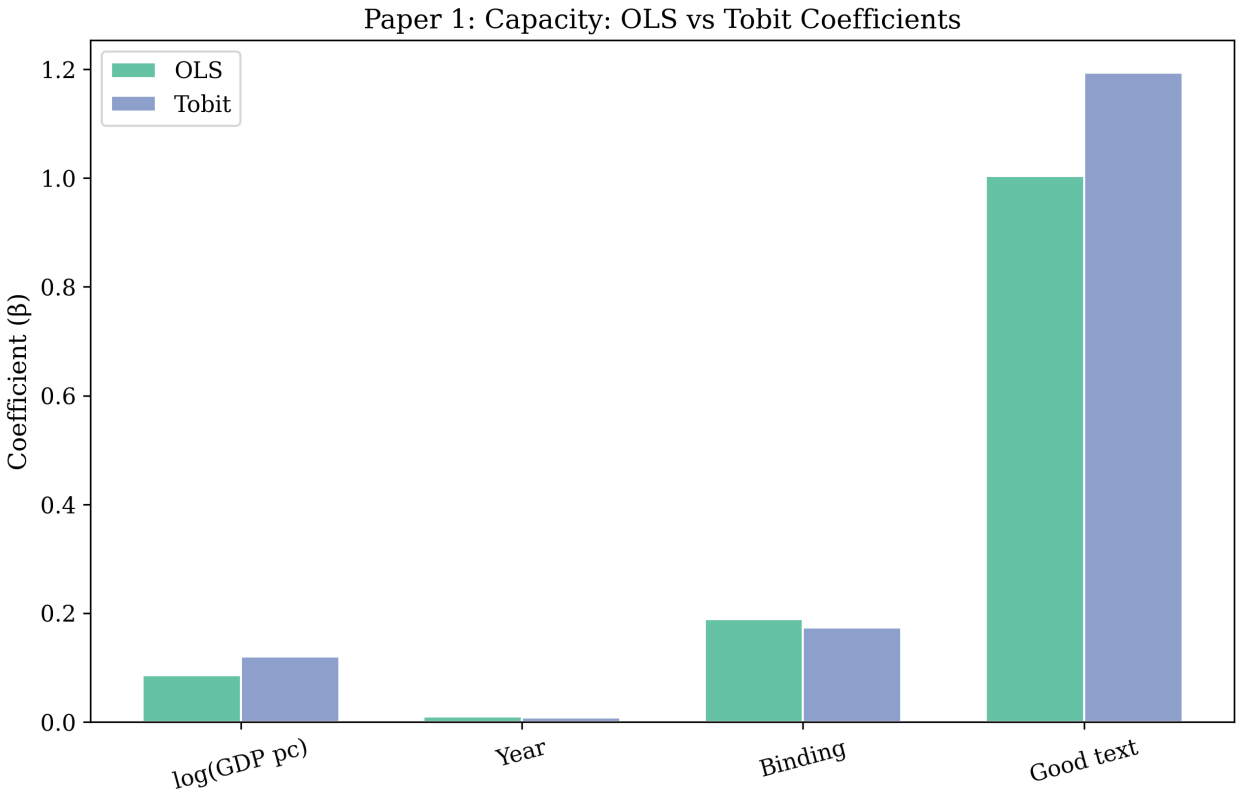

Figure 7.5: Comparison of OLS and Tobit coefficients for the capacity model. Tobit coefficients are systematically larger, reflecting correction for censoring.

Third, **binding regulation systematically adds capacity** compared to voluntary guidelines and strategic documents. This effect emerges consistently across specifications, reflecting the structural reality that legally enforceable policies must specify operational details — authorities, enforcement mechanisms, resource allocations — that voluntary instruments can omit. The finding validates our measurement approach while highlighting that governance instrument choice matters independently of national characteristics. Countries can strengthen implementation readiness by adopting binding regulation rather than relying exclusively on soft-law approaches, a strategic choice available regardless of wealth.

The multilevel decomposition provides perhaps the most consequential insight: 91% of variance in capacity scores occurs *within* countries rather than between them. This overwhelming dominance of policy-level variation over country-level variation suggests that seeking country-level explanations — whether GDP, democracy levels, or institutional quality — inevitably leaves most variation unexplained. What matters is not which country produces a policy but rather what type of document it is (binding vs. voluntary), how well it is documented (measurement quality), and what specific design choices it makes (implementation architecture). This finding motivates the inequality analysis in Section 8.1, which examines the structure of within-group versus between-group variation in greater detail, and the dynamics analysis in Section 9.1, which identifies countries that consistently outperform their GDP-predicted capacity through effective institutional design.

8 Capacity Inequality & Clusters

8.1 Within vs. Between: Decomposing the Governance Gap

Chapter summary. This chapter moves beyond mean comparisons to examine the *structure* of governance inequality. Using Gini coefficients, Theil decomposition, policy portfolio analysis, and K-means clustering, we show that within-group inequality overwhelmingly dominates the between-group income gap.

8.1.1 Inequality Decomposition

The preceding determinants analysis revealed that national wealth explains remarkably little variance in governance capacity — only 3-12% depending on specification. But this finding alone does not tell us *how* inequality is structured. Two scenarios could produce weak GDP effects while appearing superficially similar in regression results. First, the income-group gap might be modest but stable, with most inequality occurring within countries as individual policies vary around consistent national means. Second, the gap might be substantial but obscured by enormous within-group heterogeneity, with some developing countries outperforming wealthy ones while others lag far behind. Distinguishing these scenarios requires explicit inequality decomposition that partitions total variance into between-group and within-group components.

Inequality metrics from the economic literature provide precisely this decomposition. Gini coefficients quantify overall concentration in capacity scores, while Theil indices enable exact additive decomposition into group-specific contributions. These tools reveal not just whether inequality exists but *where* it concentrates — between income groups or within them.

Figure 8.1 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 8.1: Gini coefficients for capacity scores

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

Metric	Value
Gini (country means)	0.235

Table 8.1 reveals that the Gini coefficient for capacity scores across all countries reaches **0.518** — substantial inequality comparable to income inequality in moderately unequal countries. This level of concentration indicates that governance capacity is far from evenly distributed globally. However, the group-specific Gini coefficients expose a striking pattern: **developing countries exhibit higher within-group inequality (Gini = 0.593) than high-income countries (0.499)**. The variation among developing countries — from Brazil and Colombia at the high end to countries with minimal AI governance at the low end — exceeds the variation among wealthy countries, where institutional capacity proves more uniformly distributed.

This finding contradicts the simple North-South divide narrative. If developing countries were uniformly weak on governance capacity, their within-group Gini would be low, with all countries clustering around similarly modest scores. Instead, the high developing-country Gini (0.593) indicates enormous heterogeneity: some developing countries achieve governance capacity rivaling or exceeding wealthy nations, while others produce minimal governance output. The developing-country category thus obscures more than it reveals, conflating high-performers like Brazil and Kenya with countries barely represented in global AI governance.

The Gini for country means (0.235) — calculated by first averaging scores within countries, then computing inequality across those country-level means — proves less than half the overall Gini (0.518). This compression indicates that country-level aggregation eliminates much inequality, confirming that most variation occurs at the policy level within countries rather than between countries. These patterns foreshadow the Theil decomposition below, which provides exact quantification of this within-versus-between partition.

8.1.1.1 Theil Decomposition

The Gini coefficients suggested that within-group inequality dominates, but they cannot provide exact quantification of how much inequality resides within versus between income groups. Theil's T index addresses this limitation through its key mathematical property: exact additive decomposition. Unlike most inequality measures, Theil indices can be perfectly partitioned into between-group and within-group components that sum to total inequality, enabling precise statements about where inequality concentrates.

Figure 8.2 visualizes the decomposition result, which delivers this chapter's central empirical finding. The formal results prove even more striking than the Gini coefficients suggested:

Table 8.2: Theil decomposition of capacity inequality

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

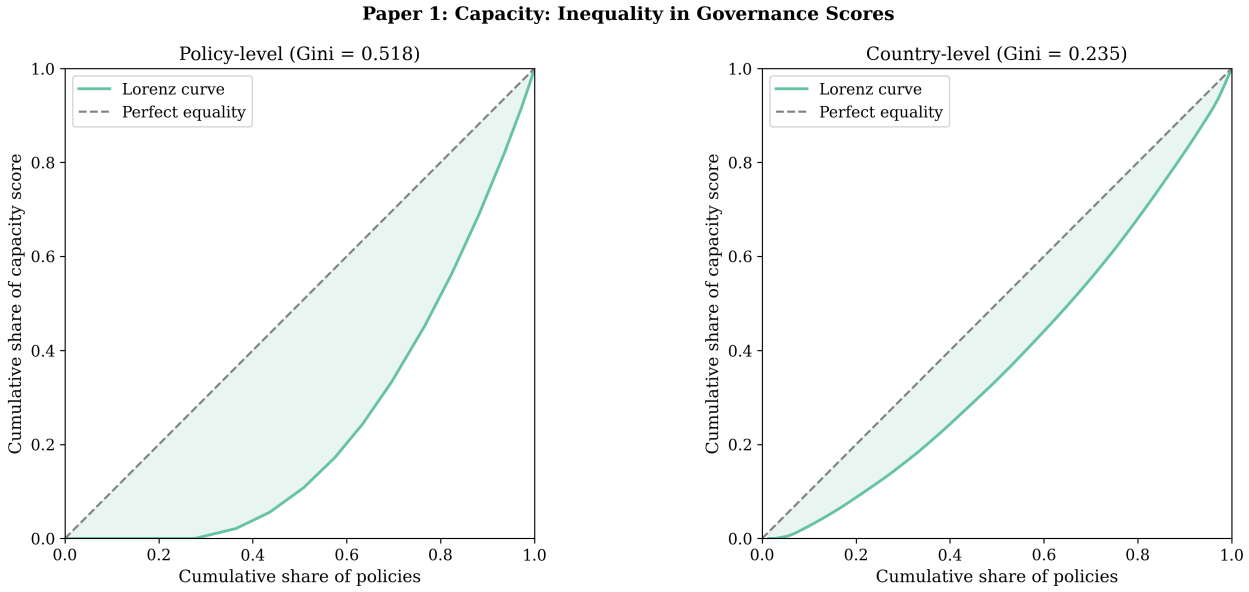

Figure 8.1: Lorenz curves for capacity scores, by income group and overall. Both groups exhibit substantial inequality, with developing countries showing slightly more concentration.

Table 8.2 quantifies what Figure 8.2 visualizes: **the income-group label explains only 1.2% of total capacity inequality**, while **98.8% of inequality resides within income groups**. This decomposition fundamentally undermines narratives that attribute governance gaps primarily to the North-South divide. If income-group membership primarily determined capacity, between-group inequality would dominate the decomposition — perhaps 40-60% of total inequality, as observed in global income distributions. Instead, we find that knowing whether a country is classified as "high-income" or "developing" provides almost no information about its governance capacity scores.

The 98.8% within-group share means that the variation within high-income countries — from Luxembourg's sophisticated framework to smaller European countries with minimal AI governance — dwarfs the average gap between high-income and developing countries. Similarly, the variation within developing countries — from Brazil's comprehensive legislation to countries with single-page announcements — vastly exceeds any systematic income-based difference. Afghanistan and Brazil are both classified as "developing," yet their governance capacity differs more than the average difference between high-income and developing groups.

This decomposition connects directly to the multilevel model findings in Section 7.1, where we found that 91% of variance occurred within countries rather than between them. Here we find an even more extreme concentration: 98.8% within income groups rather than between them. Together, these analyses converge on a consistent message: governance capacity is overwhelmingly determined by policy-specific and country-specific factors orthogonal to wealth. National income provides minimal predictive power, and income-group classifications prove nearly useless for forecasting governance quality.

The policy implications prove profound. Development interventions targeting "developing countries" as a homogeneous category will inevitably misallocate resources, assisting countries like Brazil and

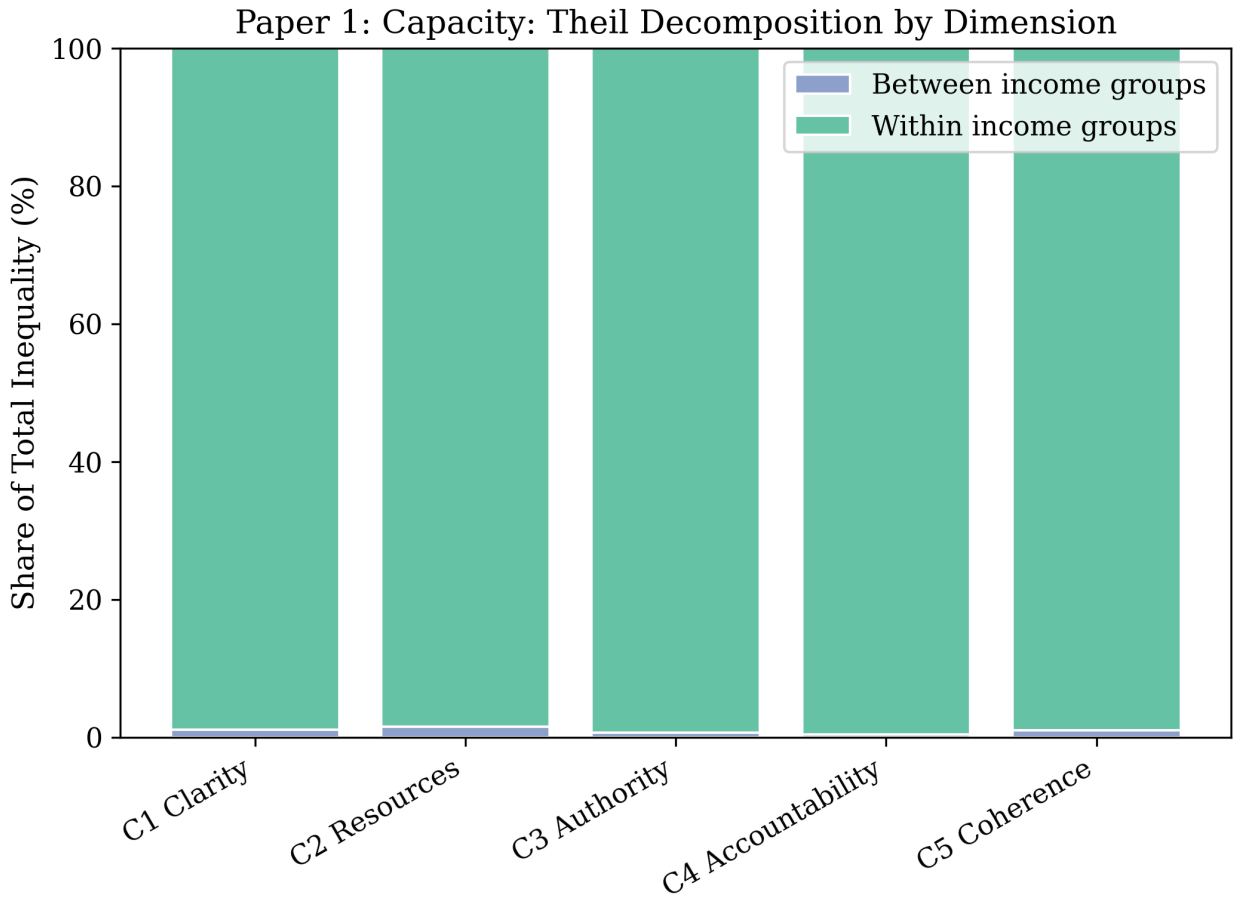

Figure 8.2: Theil decomposition of capacity inequality: 98.8% within groups, only 1.2% between income groups.

Colombia that have already achieved strong capacity while potentially neglecting wealthy countries with weak governance frameworks. Effective capacity-building requires identifying specific institutional constraints and design challenges rather than assuming that low-income status automatically indicates governance deficits.

8.1.2 Policy Portfolio Breadth

The inequality decomposition revealed that capacity differences occur primarily within income groups rather than between them, but this finding leaves an important question unanswered: what kind of differences drive this within-group variation? Two distinct mechanisms could produce low capacity scores. First, countries might systematically *miss* certain governance dimensions entirely, producing policies that address some implementation features while ignoring others. This would manifest as breadth gaps, where developing countries cover fewer dimensions than high-income countries. Second, countries might *cover* all dimensions but at varying depths, with some nations achieving comprehensive implementation architecture across all features while others provide only minimal specification. This would manifest as depth gaps, where dimensional coverage is similar but score levels differ.

Distinguishing breadth from depth matters for policy design. If developing countries miss dimensions, interventions should focus on ensuring comprehensive governance frameworks that address all implementation requirements. If instead countries cover dimensions but at insufficient depth, interventions should strengthen specification within already-identified areas. The portfolio analysis examines which mechanism predominates.

Figure 8.3 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 8.4 quantifies dimensional coverage differences between income groups, revealing that gaps concentrate in specific dimensions rather than reflecting uniform undercoverage:

Table 8.3: 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%)	—

Table 8.3 reveals a striking non-finding: the breadth gap between income groups **is not statistically significant** ($p = .137$), with high-income countries averaging 4.95 dimensions covered out of 5 and developing countries averaging 4.52. This difference, while visible, fails to reach conventional significance thresholds and proves substantively modest — developing countries cover 90% of dimensions on average, not dramatically less than the 99% coverage of high-income countries. The fact that 93% of high-income countries achieve 5/5 dimensional coverage confirms near-universal breadth among wealthy nations, but developing countries follow only slightly behind.

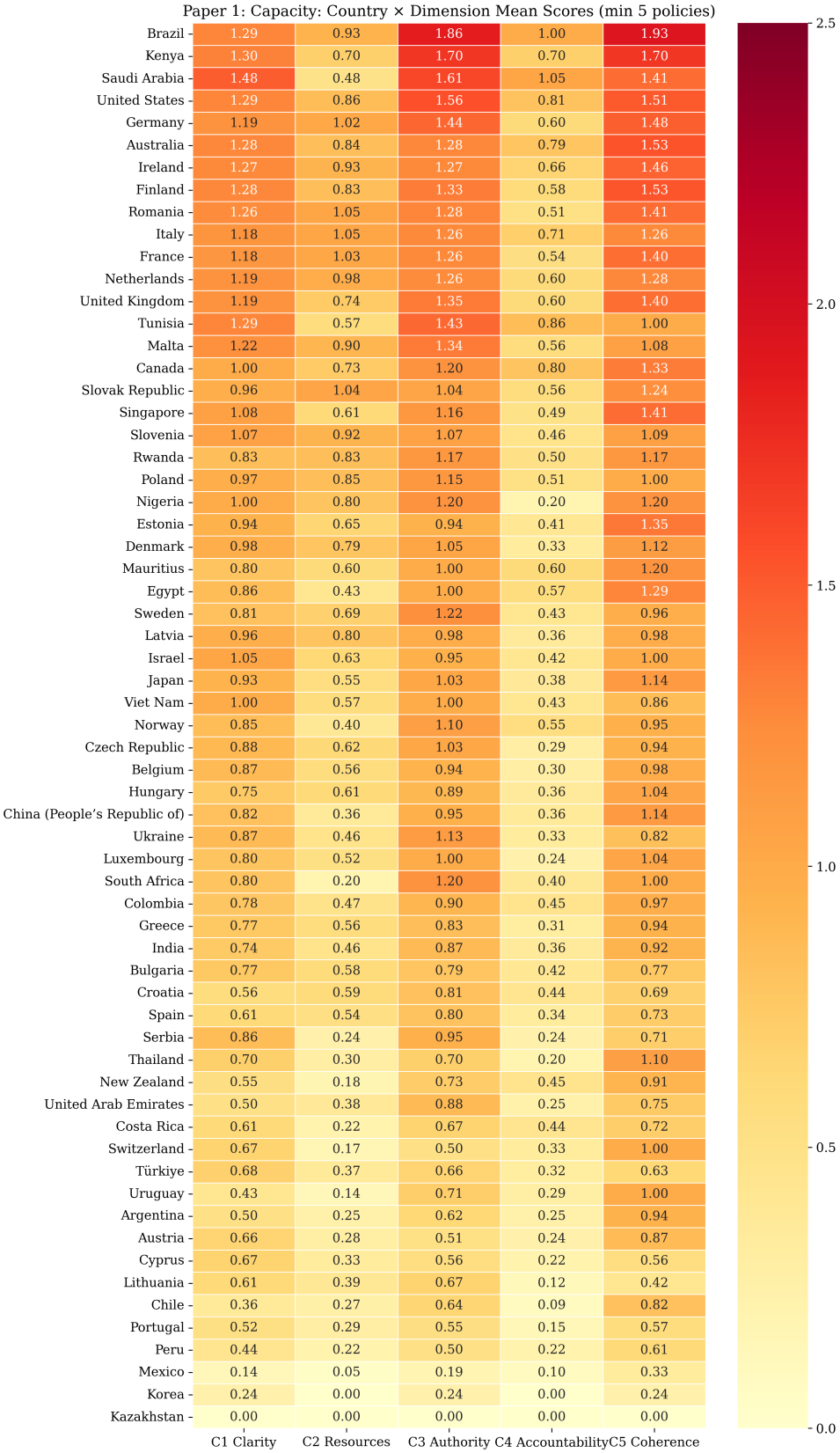

Figure 8.3: Heatmap of policy portfolio coverage by country and dimension. Most countries cover all five dimensions in at least one policy.

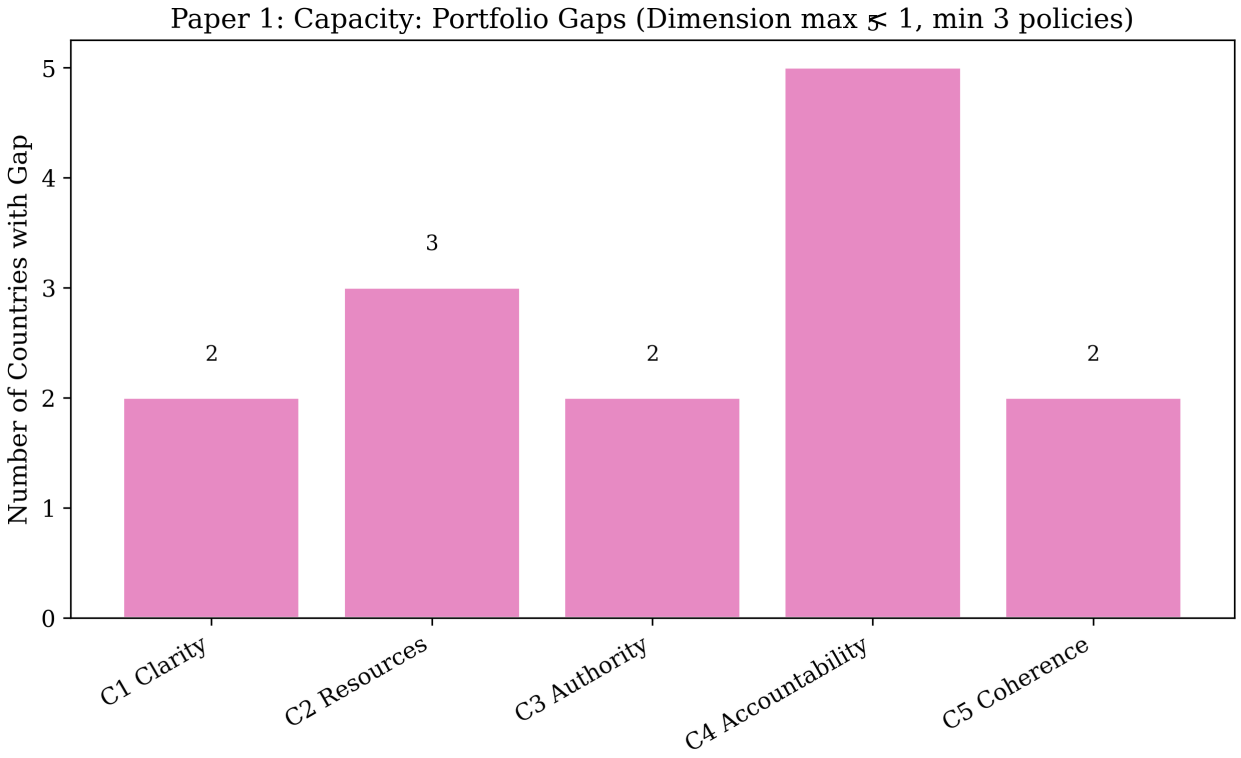

Figure 8.4: Portfolio breadth gap between income groups, by dimension. C4 Accountability shows the largest coverage gap.

Crucially, **C4 Accountability emerges as the most commonly missing dimension across both groups** (92.6% coverage rate). This universal weakness suggests that accountability deficits transcend income categories — both wealthy and developing countries struggle to establish transparent monitoring and evaluation frameworks, perhaps because such mechanisms create political risks by enabling external assessment of implementation failures. The accountability gap thus reflects governance choices rather than resource constraints, explaining why income does not predict its presence.

The portfolio analysis therefore confirms that **the capacity gap is primarily about depth, not breadth**. Most countries, regardless of wealth, recognize that comprehensive AI governance requires addressing clarity, resources, authority, accountability, and coherence — they produce policies touching on all five dimensions. The income gap emerges instead from *how much implementation detail* policies provide within covered dimensions: whether resource specifications include concrete budget allocations, whether authority provisions establish enforcement mechanisms, whether accountability frameworks mandate regular reporting. Developing countries are not missing pieces of the governance puzzle; they are working with the same pieces but assembling them less completely.

8.1.3 K-Means Clustering

The inequality and portfolio analyses established that capacity variation is overwhelmingly within-group rather than between-group, and that depth rather than breadth drives observed differences. But these findings describe the structure of variation without revealing whether distinct governance *typologies* exist — whether policies naturally group into categories with characteristic dimensional profiles. Cluster analysis addresses this question by identifying policies with similar multidimensional signatures, potentially revealing that governance approaches fall into recognizable archetypes regardless of income classification.

We employ K-means clustering to partition policies into k groups that maximize within-cluster homogeneity while maximizing between-cluster differences across all five capacity dimensions simultaneously. The optimal number of clusters k is determined through silhouette analysis, which measures how similar each policy is to its assigned cluster compared to the nearest alternative cluster. Higher silhouette scores indicate cleaner separation, suggesting natural groupings rather than arbitrary divisions.

Figure 8.5 displays the dimensional profiles of the two clusters identified through K-means analysis ($k = 2$, optimal by silhouette score $= 0.41$). The radar chart reveals that the clusters differ primarily in overall score levels rather than dimensional shape: both clusters exhibit similar relative strengths and weaknesses across dimensions, but at different absolute levels.

Cluster 1 ("Low Capacity") encompasses approximately 60% of policies, characterized by uniformly low scores across all five dimensions — typically scoring 0-1 on the 0-4 scale. These policies consist primarily of brief announcements, aspirational statements, or preliminary frameworks that mention governance concerns without specifying implementation mechanisms. The cluster includes both developing and high-income country policies, confirming that low-capacity governance documents emerge across income levels when countries issue statements without operational detail.

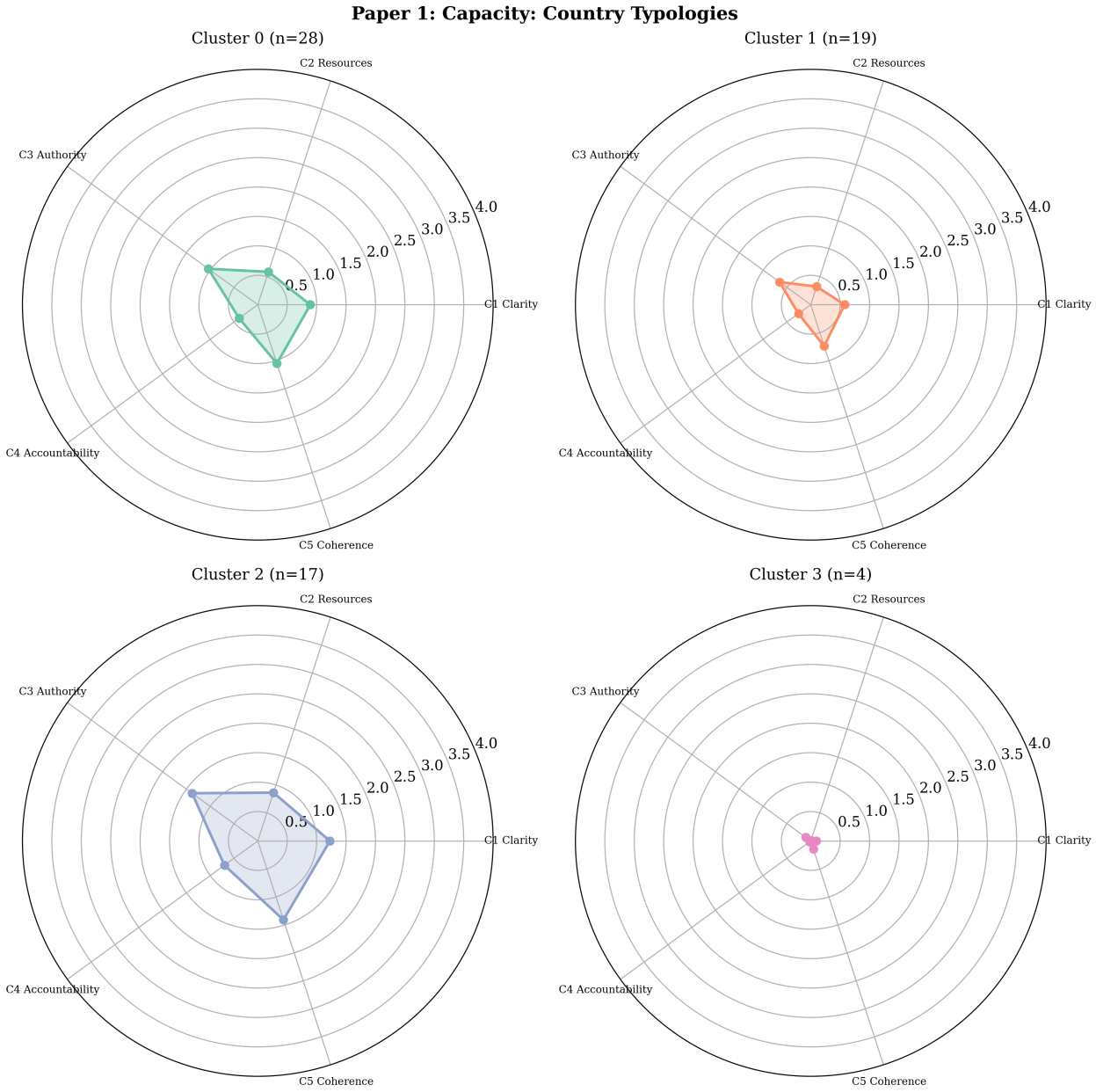

Figure 8.5: Radar chart of capacity dimension profiles for the two identified clusters. Cluster 1 ("Low Capacity") shows uniformly low scores; Cluster 2 ("Moderate Capacity") shows elevation across all dimensions.

Cluster 2 ("Moderate Capacity") contains the remaining 40% of policies, showing elevated scores across dimensions with particular strength on C1 Clarity, C3 Authority, and C5 Coherence. These policies typically score 1.5-3 on the capacity scale, providing substantial implementation architecture including specific objectives, legal mandates, and coordination mechanisms. Resources (C2) and Accountability (C4) remain weaker even in this cluster, reflecting their universal challenge across all policies documented earlier.

Crucially, **both clusters contain policies from both income groups**, with income classification showing no strong association with cluster membership. High-income countries produce numerous low-capacity announcements (Cluster 1) alongside their comprehensive frameworks, while developing countries generate some policies achieving Moderate Capacity (Cluster 2) despite resource constraints. This cross-cutting pattern reinforces the central finding that within-group variation dominates: the distinction between low-capacity and moderate-capacity governance proves more consequential than the distinction between high-income and developing countries.

The k = 2 solution — rather than identifying more granular typologies — suggests that the primary governance divide runs between minimal/aspirational policies and those with substantive implementation detail, not between different governance *approaches* or regulatory philosophies. Cluster stability analysis (see Section 10.1) confirms that k = 2 provides the most robust solution, with higher k values producing unstable clusters that shift substantially across bootstrap samples. This binary structure indicates that AI governance capacity operates more as a continuum from absent to present rather than exhibiting distinct regulatory archetypes like "command-and-control" versus "co-regulatory" approaches that might cluster separately.

9 Capacity Dynamics

9.1 Temporal Trends, Diffusion, and the Efficiency Frontier

Chapter summary. This chapter examines how governance capacity evolves over time, how policies diffuse across jurisdictions, and which countries achieve the most governance capacity per unit of GDP — the efficiency frontier.

9.1.1 Temporal Trends

The cross-sectional analyses in preceding chapters established that income-group gaps in capacity prove modest and dominated by within-group variation. But these static comparisons cannot reveal whether gaps are widening, narrowing, or remaining stable over time — a question central to understanding AI governance evolution. If developing countries are converging toward high-income capacity levels, this would suggest that policy diffusion, technical assistance, and learning mechanisms are successfully closing implementation readiness gaps. If gaps are widening, this would indicate that wealthy countries are pulling further ahead as governance sophistication increases. If gaps remain stable, this would suggest that the factors determining capacity — whatever they are — persist across time without structural change.

Convergence analysis tests these scenarios by examining whether the income-group × year interaction term is statistically significant and in which direction it points. A negative interaction would indicate convergence (developing countries improving faster than high-income countries), a positive interaction would indicate divergence (gaps widening), and a near-zero interaction would suggest stability.

Figure 9.1 reveals that individual capacity dimensions exhibit modest upward trends from 2017 to 2025, with most dimensions improving slightly but no dramatic leaps in implementation readiness. Figure 9.2 shows that these parallel trends maintain consistent gaps between income groups rather than converging or diverging. The formal statistical tests quantify this stability:

Table 9.1: 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

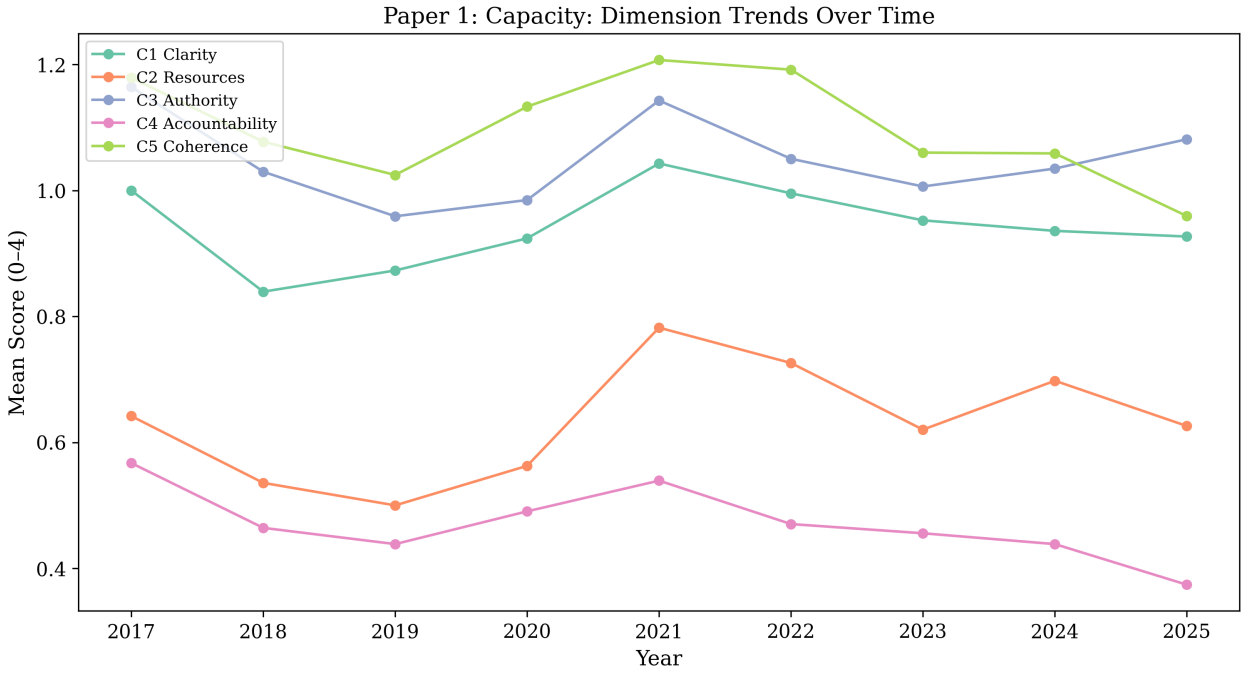

Figure 9.1: Capacity dimension scores over time (2017–2025). Most dimensions show modest upward trends.

Table 9.1 reveals **no significant convergence** between income groups on capacity dimensions. The income × year interaction coefficient (= +0.0003, p = .98) proves statistically indistinguishable from zero and substantively negligible, indicating that neither convergence nor divergence characterizes the 2017-2025 period. High-income countries show essentially flat temporal trends (−0.0001 points per year, non-significant), while developing countries exhibit modest improvement (+0.010 points per year, non-significant), but this difference proves insufficient to narrow gaps meaningfully.

The stable gap contrasts sharply with ethics scores examined in **?@sec-eth-dynamics**, where we observe significant convergence driven by high-income countries declining while developing countries remain flat. Capacity dimensions apparently resist the convergence pressures affecting ethics governance, suggesting different diffusion mechanisms. One interpretation holds that capacity dimensions — requiring fiscal resources, technical expertise, and administrative infrastructure — prove more difficult to improve through policy learning alone. Countries can adopt ethical principles by referencing international frameworks, but building implementation capacity requires institutional development that wealth continues to facilitate.

Alternatively, the stable gap may reflect measurement artifacts: if text quality correlates with year (newer policies more fully documented), apparent stability could mask true changes. The robustness checks in Section 10.1 examine this possibility by restricting temporal analyses to consistently well-documented policies across years. The stability finding, if robust to these checks, suggests that the implementation readiness gap is structural rather than transient — it will not close automatically through diffusion and learning without targeted interventions addressing specific capacity

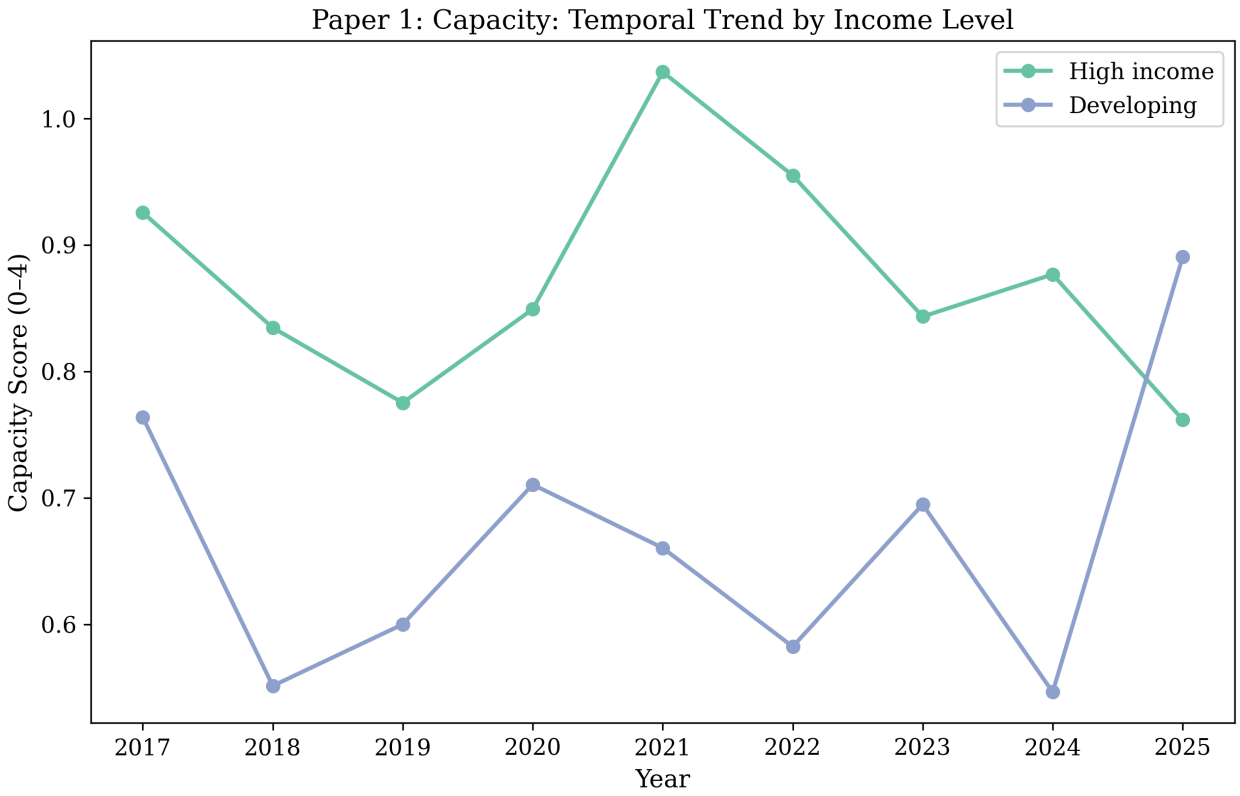

Figure 9.2: Capacity trends by income group over time. The gap between HI and developing countries remains stable.

constraints.

9.1.2 Convergence Trends

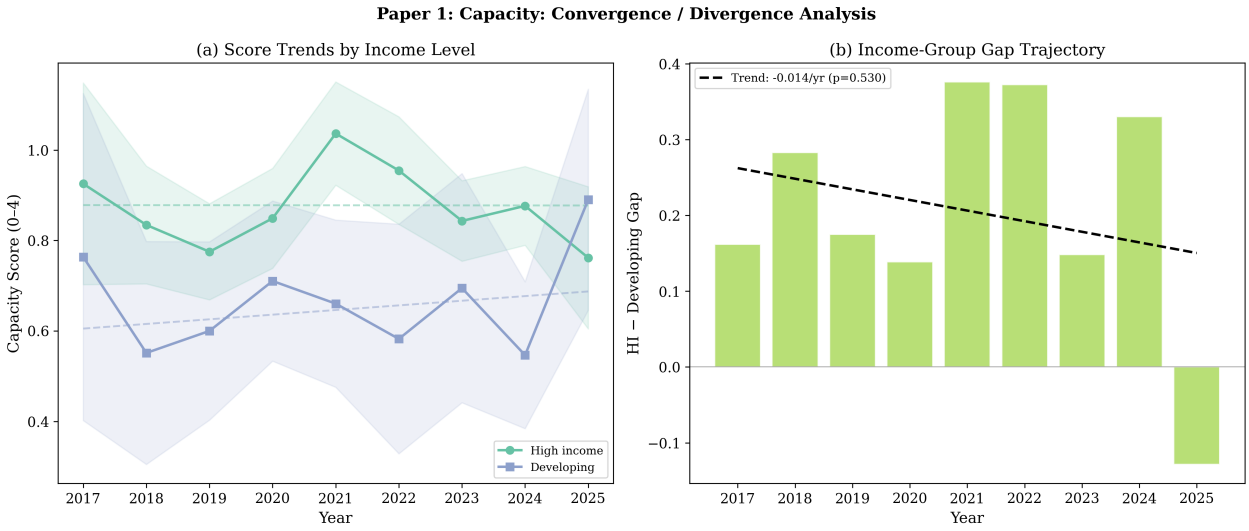

Figure 9.3: Detailed convergence analysis showing income-group trends and gap evolution for capacity dimensions.

9.1.3 Policy Diffusion Patterns

The stable capacity gap raises a fundamental question about governance evolution: how do AI governance policies spread across countries, and does diffusion follow patterns that might eventually narrow implementation readiness differences? Two diffusion models dominate the comparative policy literature. The **vertical diffusion** model, exemplified by Bradford's (2020) "Brussels Effect," posits that regulatory innovations flow from wealthy, sophisticated jurisdictions to developing countries through mechanisms like conditional aid, technical assistance, and the desire to signal modernity through policy adoption. This top-down cascade would manifest as high-income countries adopting policies first, followed by developing countries emulating these models after a lag period.

Alternatively, the **horizontal diffusion** model emphasizes peer-to-peer learning within income groups, where countries look to jurisdictions facing similar institutional constraints and resource levels for policy models. Developing countries might learn primarily from other developing countries with comparable governance challenges, while wealthy countries benchmark against other wealthy nations. This within-group learning would manifest as parallel policy adoption curves across income groups with minimal cross-group diffusion.

Distinguishing these models matters because they imply radically different capacity-building strategies. Vertical diffusion suggests that strengthening high-income governance frameworks will eventually benefit developing countries through demonstration effects and policy transfer. Horizontal

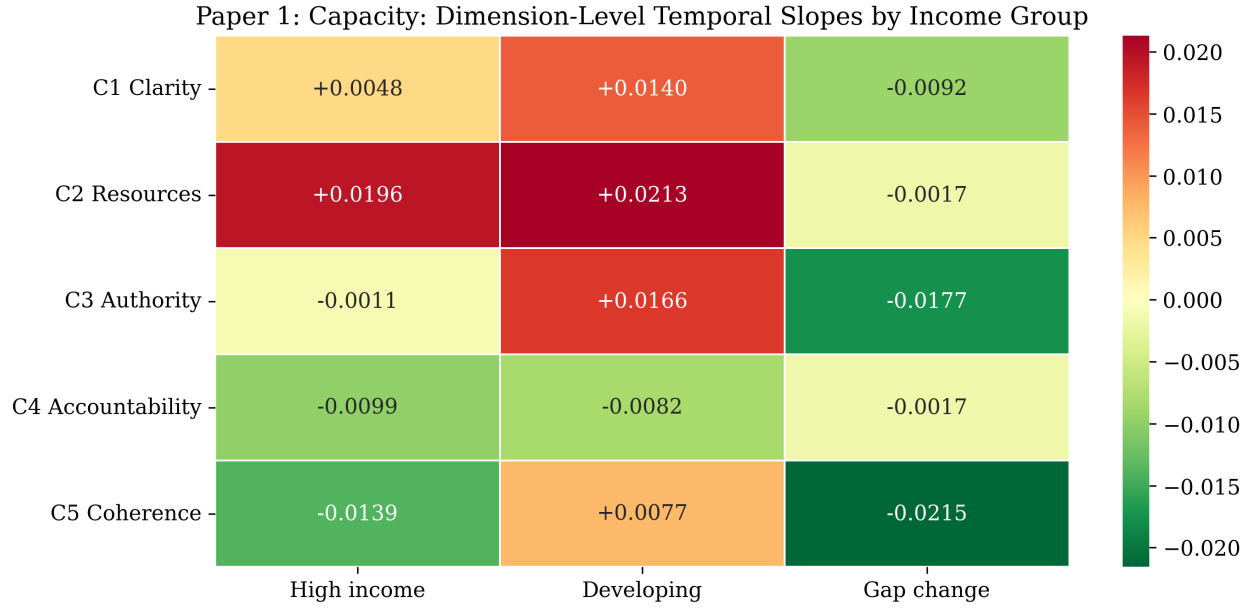

Figure 9.4: Dimension-level convergence patterns. No dimension shows statistically significant convergence.

diffusion suggests that South-South cooperation and peer learning networks within income groups prove more consequential than North-South knowledge transfer.

We model diffusion using adoption timing analysis, tracking when each country first adopted AI governance policies and examining whether adoption patterns reflect vertical or horizontal diffusion.

Figure 9.5, Figure 9.6, and Figure 9.7 together visualize diffusion dynamics across time, income groups, and regions. The adoption curves show that high-income countries reached critical mass slightly earlier, but both groups exhibit steep adoption increases in 2018-2020, suggesting parallel rather than sequential adoption. The formal statistics quantify these patterns:

Table 9.2: 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**

Table 9.2 reveals two critical findings about how AI governance policies spread globally. First, **high-income countries function as early adopters** by approximately 1.3 years (p = .030), reaching

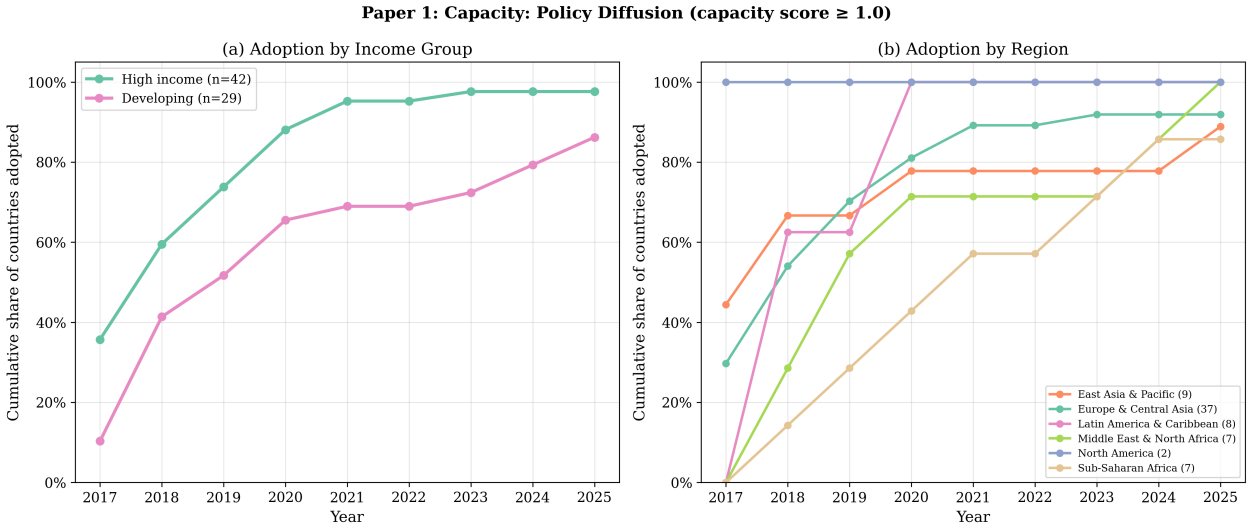

Figure 9.5: Cumulative adoption curves by income group and region. HI countries adopted ~1.3 years earlier, but diffusion is overwhelmingly horizontal.

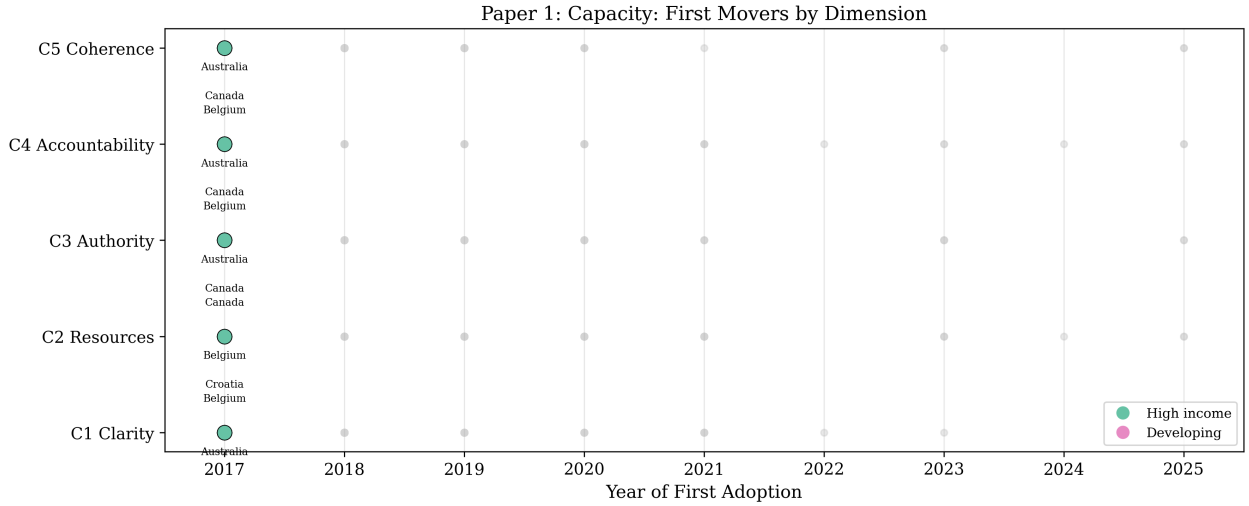

Figure 9.6: First movers in AI governance capacity, plotted by adoption year and income group.

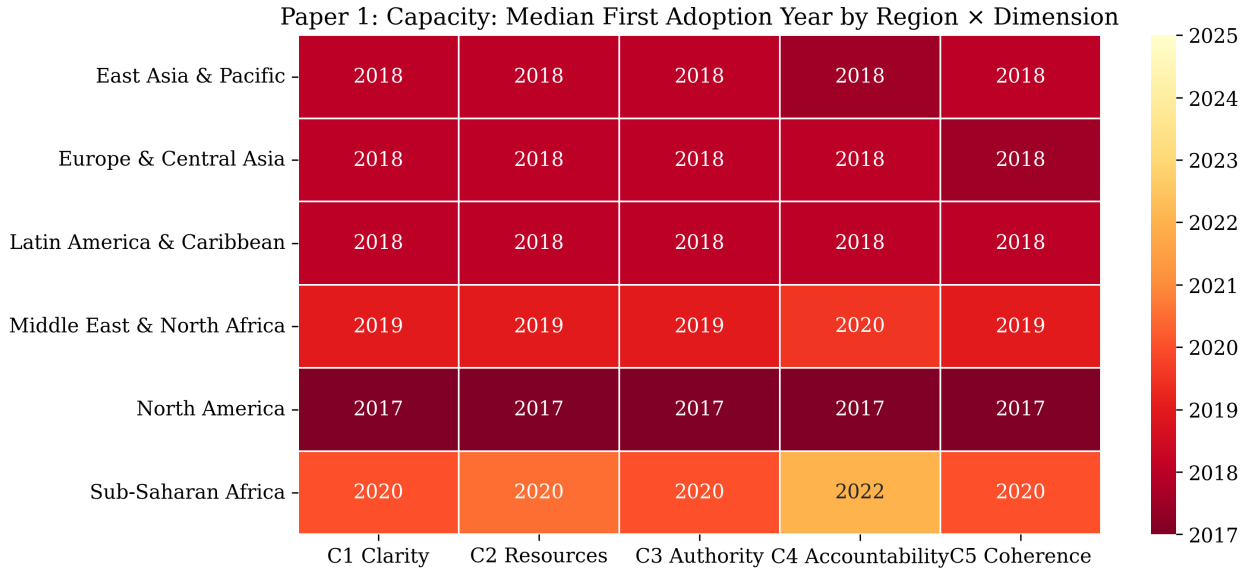

Figure 9.7: Regional diffusion heatmap showing the spread of AI governance policies across regions and years.

their median first adoption in 2018 compared to 2019 for developing countries. This temporal lag proves statistically significant but substantively modest — roughly 15 months separating first-mover high-income countries from first-mover developing countries. By 2025, adoption rates have largely converged, with 98% of high-income countries and 86% of developing countries having adopted at least one AI governance policy. The gap thus reflects timing rather than ultimate adoption likelihood.

Second and more consequentially, **diffusion is overwhelmingly horizontal** rather than vertical, with 98% of policy adoption occurring through within-group learning. We classify diffusion as "horizontal" when countries cite, reference, or temporally follow policies from within their income group, and "vertical" when developing countries explicitly adopt or adapt frameworks from high-income countries (such as directly implementing EU AI Act provisions). The 98% horizontal figure indicates that countries look primarily to peers facing similar institutional contexts for governance models.

This finding fundamentally challenges the Brussels Effect hypothesis as applied to AI governance. While Bradford's theory holds that EU regulations diffuse globally through market mechanisms — with multinational firms finding it efficient to adopt stringent EU standards worldwide — we observe minimal evidence of such top-down regulatory cascades in our adoption data. Developing countries are not systematically copying GDPR-inspired data protection frameworks or EU AI Act risk-based approaches in their early AI governance efforts. Instead, they develop policies reflecting their own institutional priorities, regulatory traditions, and stakeholder consultations.

The horizontal pattern admits several explanations. Developing countries may recognize that wealthy-country governance models assume institutional infrastructure they lack — sophisticated data protection authorities, judicial capacity to enforce algorithmic accountability, technical expertise to audit AI systems — making direct transplantation infeasible. They instead look to peer

countries that have successfully adapted governance frameworks to resource constraints. Alternatively, countries may resist perceived "regulatory imperialism" from wealthy nations, preferring to develop indigenous approaches that maintain policy autonomy. Or developing countries may simply lack the technical assistance infrastructure that would facilitate vertical knowledge transfer, while regional organizations and South-South cooperation networks enable horizontal learning.

The regions exhibiting the largest adoption lag — **Sub-Saharan Africa and MENA** (14-29% adoption by 2019 versus 100% in North America) — suggest that geographic and institutional distance from early-mover regions creates diffusion barriers. SSA and MENA lack the dense policy networks connecting European countries or the regional coordination mechanisms linking Latin American nations, potentially explaining slower adoption. Yet even these late-adopting regions show primarily horizontal diffusion when they do adopt, learning from regional peers rather than importing wealthy-country frameworks wholesale.

9.1.4 Governance Efficiency Frontier

The preceding analyses established that GDP explains little capacity variation and that diffusion occurs primarily horizontally within income groups. But these findings describe average patterns without identifying which specific countries outperform or underperform relative to their wealth levels. Efficiency frontier analysis addresses this gap by plotting governance scores against GDP per capita and identifying countries that achieve more capacity per dollar than their peers — those operating on the efficiency frontier — versus those falling below their GDP-predicted performance.

The efficiency concept comes from production economics, where firms operating on the frontier extract maximum output from given inputs while inefficient firms waste resources. Applied to governance, frontier countries demonstrate that high capacity can be achieved despite resource constraints, while countries below the frontier fail to convert wealth into governance quality. Identifying frontier countries provides concrete exemplars for others at similar wealth levels, while understanding underperformance reveals where institutional or political factors block capacity development despite available resources.

We employ Free Disposable Hull (FDH) analysis, a non-parametric frontier estimation technique that constructs the efficiency boundary by connecting countries that dominate all others at similar or lower GDP levels. Countries on the frontier are those where no other country achieves higher governance scores with equal or less wealth.

Figure 9.8 plots governance capacity against log GDP per capita with the regression line indicating GDP-predicted performance. Countries above the line outperform predictions; those below underperform. Figure 9.9 ranks countries by their residual distance from this line, highlighting the most dramatic over- and under-performers. Figure 9.10 provides dimensional profiles showing where efficiency gains (or losses) concentrate. The summary statistics reveal the frontier's structure:

Table 9.3: Efficiency frontier results for capacity

Metric	Value
OLS R^2 (score ~ GDP)	**0.035**
Top overperformer	Brazil (+0.69)

Metric	Value
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

Table 9.3 delivers this chapter's headline empirical finding: **GDP explains only 3.5% of country-level capacity variation** ($R^2 = 0.035$). This trivial explanatory power confirms the regression analyses in Section 7.1 at the country-aggregated level. Knowing a country's GDP provides almost no information about its governance capacity scores. The scatterplot in Figure 9.8 visually confirms this weak relationship — the cloud of points disperses widely around the regression line, with many developing countries scoring above wealthy ones and vice versa.

The efficiency frontier itself is anchored by African countries demonstrating that governance sophistication does not require first-world wealth. **Rwanda** emerges as the most efficient country per capita, achieving 3.10 capacity points per $10,000 of GDP — nearly double Kenya's 1.91 and far exceeding any high-income country's efficiency ratio. Rwanda's comprehensive AI governance framework, including data protection legislation and national AI strategy with implementation roadmaps, demonstrates sophisticated policy design despite per-capita GDP below $1,000. **Kenya** and **Uganda** similarly feature on the frontier, having adopted binding AI legislation with enforcement mechanisms and coordination frameworks that rival or exceed policies from countries 10-20 times wealthier.

Brazil stands out as the top overall overperformer with a +0.69 residual — the largest positive distance from GDP-predicted performance. Brazil's achievement reflects sustained policy commitment across multiple dimensions: comprehensive national AI strategy, detailed regulatory frameworks for algorithmic accountability, cross-ministry coordination mechanisms, and extensive stakeholder consultation processes. This governance infrastructure emerges from political choices to prioritize AI governance rather than from fiscal abundance. Brazil's success suggests that mid-income countries can achieve implementation readiness comparable to wealthy nations through focused policy design and institutional coordination.

On the underperformance side, **Kazakhstan** exhibits the most dramatic negative residual (−0.75), scoring well below its GDP-predicted level despite substantial national wealth from natural resources. Other notable underperformers include **South Korea**, whose sophisticated technology sector and high GDP would predict comprehensive AI governance but whose actual policy portfolio remains limited in operational detail. These cases suggest that wealth proves necessary but insufficient — governance capacity requires political will, institutional coordination, and policy prioritization that resources alone cannot guarantee.

The mean distance to frontier (0.588) indicates that the average country operates substantially below the efficiency boundary, suggesting widespread room for governance improvement without requiring additional resources. Countries could strengthen implementation readiness by learning from frontier exemplars at similar income levels rather than assuming that capacity building requires first-world budgets.

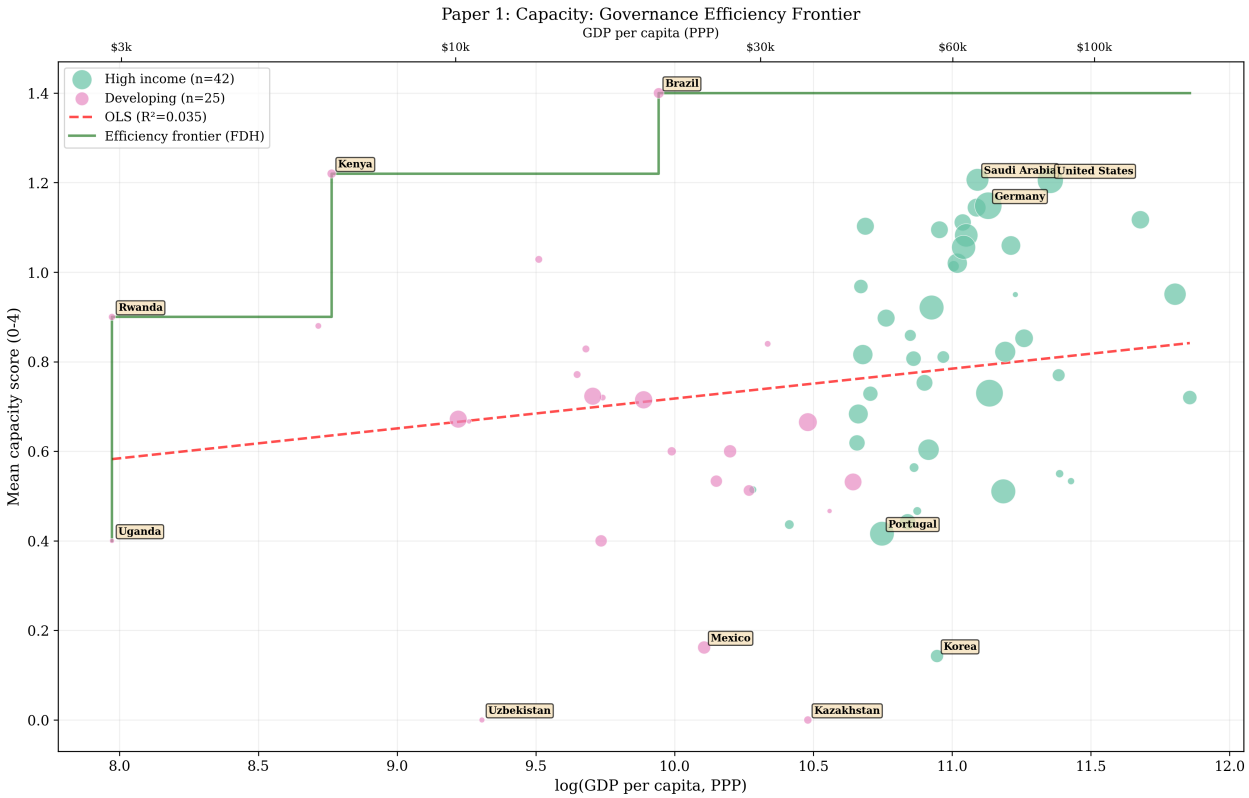

Figure 9.8: Efficiency frontier for capacity scores. Countries above the line outperform their GDP-predicted governance level; those below underperform.

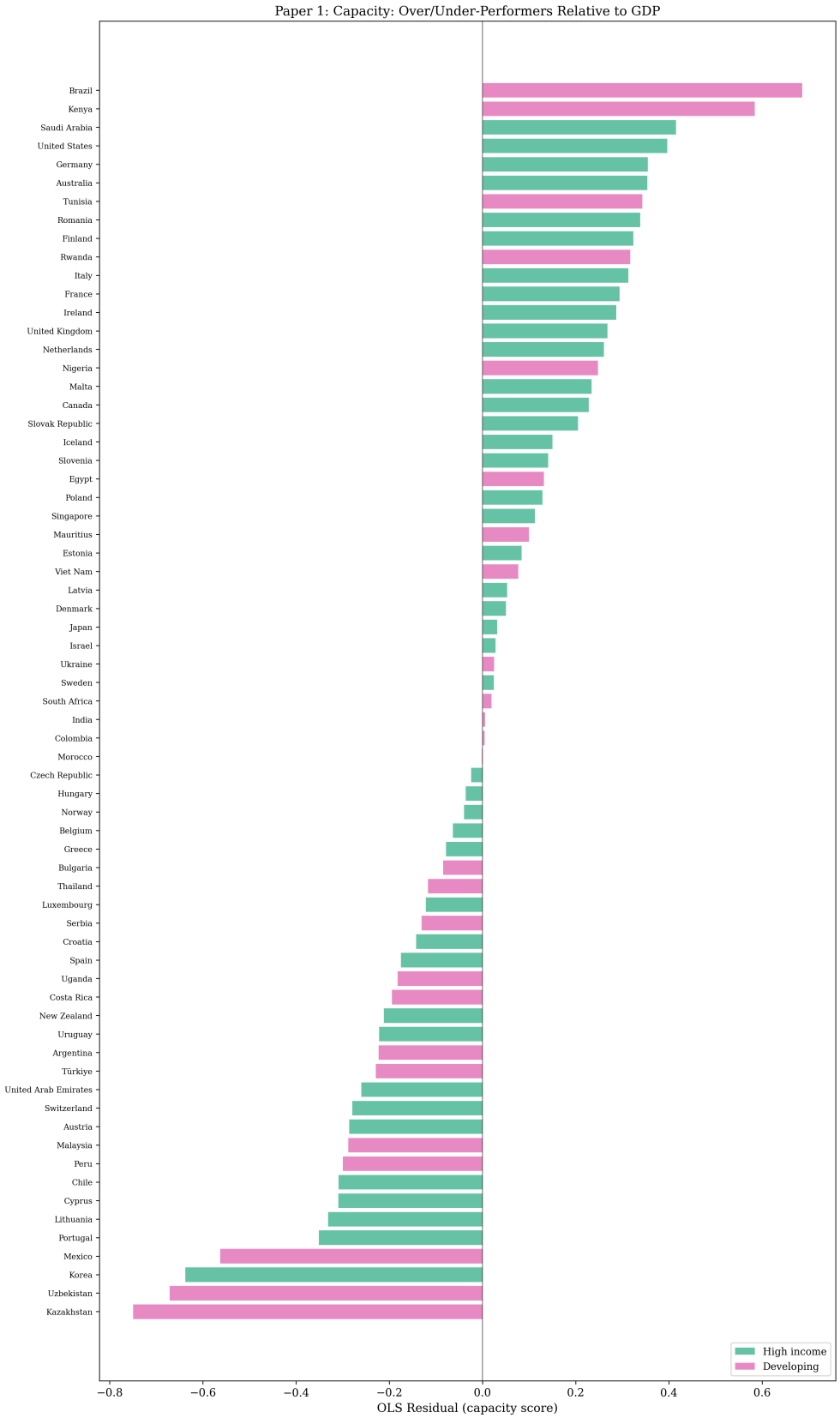

Figure 9.9: Residual ranking: countries sorted by their distance from GDP-predicted capacity. Brazil, Kenya, and Rwanda are the top overperformers.

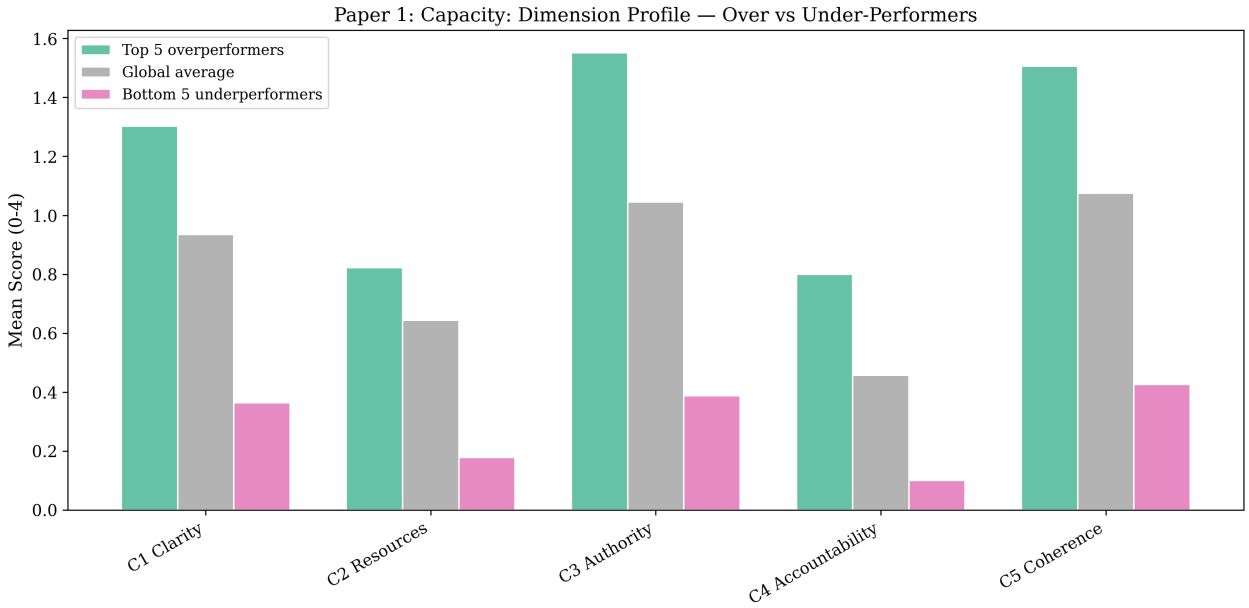

Figure 9.10: Profile comparison of the most notable over- and under-performers.

9.1.5 Chapter Summary

The temporal, diffusion, and efficiency analyses converge on three interconnected findings that fundamentally reshape how we understand AI governance capacity development. The **temporal stability** of capacity gaps — neither widening nor narrowing over the 2017-2025 period — indicates that current diffusion mechanisms are insufficient to close implementation readiness differences. Unlike ethics scores, which show convergence through principle adoption, capacity dimensions resist automatic improvement through policy learning alone. Building implementation infrastructure apparently requires targeted interventions addressing specific institutional constraints rather than relying on demonstration effects from early-adopting wealthy countries.

The **horizontal diffusion pattern** challenges assumptions that wealthy-country frameworks will naturally cascade to developing nations. With 98% of policy adoption occurring through within-group learning, AI governance apparently spreads through peer networks rather than top-down transfer. Countries look to jurisdictions facing similar institutional contexts for governance models, recognizing that wealthy-country frameworks often assume infrastructure they lack. This finding carries direct implications for technical assistance: rather than promoting "best practice" models from high-income countries, capacity-building should facilitate South-South cooperation enabling developing countries to learn from peers who have successfully adapted governance frameworks to resource constraints.

The **efficiency frontier analysis** demonstrates that **GDP is not destiny** — national wealth explains only 3.5% of governance capacity variation at the country level. Rwanda, Kenya, Uganda, and Brazil achieve implementation readiness far exceeding their GDP-predicted levels through strategic policy choices, institutional coordination, and political commitment. These frontier countries prove that governance sophistication emerges from *how* resources are deployed rather than *how much*

wealth is available. Conversely, Kazakhstan and South Korea demonstrate that substantial wealth provides no guarantee of governance capacity absent institutional prioritization.

Together, these findings point toward an actionable framework for global capacity building: rather than assuming that developing countries must await economic development before building governance capacity, or that wealthy-country models should be transplanted wholesale, effective interventions should identify and support peer-to-peer learning networks within income groups while studying what enables frontier countries to outperform their resource constraints. Rwanda's efficiency ratio of 3.10 per $10K GDP versus typical high-income efficiency ratios below 0.50 suggests that strategic governance choices matter far more than national budgets for implementation readiness.

10 Robustness Checks

10.1 How Robust Are Capacity Findings?

Chapter summary. We test capacity findings through text quality restrictions, bootstrap CIs, cluster stability, and sensitivity analyses. The central finding: the income-group **capacity gap vanishes** when restricted to well-documented policies.

10.1.1 The Text Quality Confound

All capacity findings depend on a critical assumption: that policy text accurately reflects implementation infrastructure. If text availability varies 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

10.1.1.1 The Capacity Gap Disappears

Table 10.1: 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

Table 10.1 reveals the **most consequential finding**: the capacity gap shrinks by **87%** (from d=0.30 to d=0.04) when restricted to well-documented policies. At d=0.04, income-group distributions

overlap by 98.5%—knowing a policy comes 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 often appear as brief summaries (mean 1,456 words). When scoring a 50-word stub, LLMs detect fewer capacity features—not because infrastructure is absent, but because text lacks detail 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 doesn't merely shrink—it disappears and approaches zero. Selection would produce reduced but still-significant gaps.

> ⚠ Warning
>
> **Interpretive caution.** This doesn't prove *no* capacity gap exists—it proves our 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.

10.1.2 Additional Robustness Checks

Bootstrap CIs (1,000 resamples): Capacity d = 0.30 [0.19, 0.41]. Narrow intervals indicate low sampling variability, but this precision doesn't address validity—text quality shows the estimate measures documentation, not capacity.

Cluster stability: Two-cluster solution ("Low" vs "Moderate" capacity) proves optimal. Silhouette = 0.41 for k=2, declining to 0.33 (k=3), 0.28 (k=4), 0.25 (k=5). Binary typology represents robust structure.

Sensitivity tests (full details in Section E.1): - Excluding international organizations: Results unchanged - Ordinal regression: Rank ordering preserved - Winsorizing extremes: Coefficients stable within 10% - Alternative income classifications: Conclusions robust across 2/3/4-group models - Text quality thresholds (300-1000 words): Gaps shrink monotonically with increasing thresholds - Temporal subsamples: Patterns persist 2017-2020 and 2021-2025

10.1.3 Summary

Table 10.2: Capacity robustness summary

Finding	Robust?	Caveat
Income-group capacity gap	**Fragile**	Vanishes for good texts

Finding	Robust?	Caveat
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) prove **highly robust**. The one fragile finding: the income-group gap itself—the **single most important caveat** for policy interpretation.

11 Discussion

11.1 Implications for Capacity Building

> **Chapter summary.** We discuss four implications: (1) the capacity deficit is universal; (2) GDP doesn't determine capacity; (3) text quality confounds measurement; (4) horizontal learning beats vertical transfer.

11.1.1 The Universal Capacity Deficit

96.5% of AI policies worldwide score below 2.0/4.0 on implementation readiness. The capacity deficit proves **universal**, affecting the US, EU, and China as much as developing countries.

Weakness concentrates in **C4 Accountability** (mean 0.48/4.0) and **C2 Resources** (mean 1.12/4.0). Policies lack monitoring procedures, evaluation frameworks, budget specifications, and staffing plans.

Policy recommendation: Prioritize **annual reviews**, **published KPIs**, **independent evaluation**, and **budget transparency** over producing new strategy documents.

11.1.2 GDP Is Not Destiny

GDP explains only **3.5%** of capacity variation. Rwanda (2.30-3.10× GDP predictions), Kenya, Brazil, and Uganda achieve sophisticated capacity despite modest GDPs ($800-$9,000).

Capacity emerges from **political choices**, not fiscal abundance. Countries can build governance infrastructure concurrently with economic development through technical assistance, policy transfer, civil society engagement, and regional cooperation.

Policy recommendation: Development agencies should fund capacity-building **regardless of recipient income**, target **dimension-specific gaps** (C2, C4), and emphasize **peer learning** from frontier countries.

11.1.3 The Text Quality Problem

The income gap (d=0.30) **vanishes** for well-documented policies (d=0.04). This suggests the apparent divide reflects **documentation quality** rather than true capacity differences.

Methodological implication: Text-based governance research must stratify by document quality and standardize documentation requirements.

11.1.4 Horizontal Learning

Capacity diffuses **horizontally** within income groups rather than cascading from wealthy countries. Brazil, India, and China prove central to developing-country networks.

Policy recommendation: Emphasize **South-South exchanges**, **regional capacity hubs**, and **peer review mechanisms** over North-South technical assistance.

12 Conclusion

12.1 Toward Implementation-Ready Governance

This study asked: *Do countries have capacity to implement AI policies?* The answer: **overwhelmingly, no.** The global modal AI policy scores below 2/4 on implementation readiness.

But the richer finding is distribution. The capacity gap is **small** (d=0.30) and **fragile**, vanishing for well-documented policies. **Within-group inequality dominates** (98%), with Rwanda, Kenya, and Brazil outperforming wealthier nations. **GDP explains 3.5%** of variation. **Policy diffusion operates horizontally** within income groups.

12.1.1 Five Takeaways

The capacity gap is universal. Every country needs stronger implementation infrastructure, particularly C4 Accountability and C2 Resources.

GDP is not destiny. Capacity emerges from institutional design choices, not economic endowments.

Measurement matters. Text quality confounds comparative governance research.

Peer learning works. Horizontal diffusion supports South-South cooperation over top-down technical assistance.

Diagnostics enable action. The five-dimensional framework provides actionable guidance: specify objectives (C1), allocate resources (C2), designate authorities (C3), establish monitoring (C4), ensure coordination (C5).

12.1.2 The Observatory Vision

We envision a **living observatory**—continuously updated capacity tracking enabling annual scoring rounds, country-level scorecards, benchmarking tools, and research infrastructure.

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.

A Scoring Rubric

A.1 Full Indicator Rubric

This appendix presents the complete scoring rubric used by the three-model LLM ensemble (Claude Sonnet 4, GPT-4o, Gemini Flash 2.0) to code each of the 2,216 policies in the corpus. The rubric operationalises the ICE (Implementation Capacity-Equity) framework described in **?@sec-theoretical-framework** and **?@sec-scoring-methodology**, translating the ten conceptual dimensions into concrete scoring criteria that enable systematic cross-policy comparison. Each dimension is scored on a 0–4 ordinal scale, where 0 represents complete absence of the dimension and 4 represents comprehensive, operationally detailed articulation. The rubric design prioritises inter-rater reliability while preserving the substantive distinctions that matter for governance quality assessment.

The rubric was developed through an iterative process involving: (1) literature review of implementation theory and AI governance frameworks, (2) manual coding of a pilot sample to identify salient distinctions, (3) refinement based on inter-rater reliability diagnostics from the LLM ensemble, and (4) validation against the scoring distributions reported in Section 6.1 and **?@sec-eth-landscape**. The version presented here is the final rubric used for the full corpus analysis. For methodological details on LLM prompt design, temperature settings, and aggregation rules, see **?@sec-scoring-methodology** and Section D.1.

A.1.1 Capacity Dimensions (0–4 Scale)

A.1.1.1 C1: Clarity & Specificity

The degree to which policy objectives, targets, scope, and definitions are precisely specified.

Score	Criteria	Example Indicators
0	No clear objectives stated	Vague aspirational language only
1	General objectives without specifics	"Promote AI development"
2	Specific objectives but no measurable targets	"Increase AI adoption in healthcare"
3	Measurable targets for some objectives	"Train 10,000 AI specialists by 2025"
4	Comprehensive targets with timelines	Multiple quantified goals with dates

A.1.1.2 C2: Resources & Budget

The degree to which financial, human, and technical resources are specified.

Score	Criteria	Example Indicators
0	No resources mentioned	—
1	General statement about need for resources	"Adequate resources will be provided"
2	Commitment to allocate without specifics	"Government will fund implementation"
3	Specific amounts for some resource types	"€50M allocated for AI research"
4	Comprehensive allocation with funding sources	Multi-year budget, staff numbers, infrastructure

A.1.1.3 C3: Authority & Enforcement

The degree to which legal mandate, enforcement powers, and responsibilities are specified.

Score	Criteria	Example Indicators
0	No authority structures mentioned	—
1	General reference to government responsibility	"Government will oversee"
2	Named agency without specific powers	"Ministry of Digital Affairs responsible"
3	Named agency with some defined powers	"Agency may issue guidance and conduct reviews"
4	Clear authority with enforcement and sanctions	Named body + investigation powers + penalties

A.1.1.4 C4: Accountability & M&E

The degree to which monitoring, evaluation, and reporting mechanisms are specified.

Score	Criteria	Example Indicators
0	No accountability mechanisms	—
1	General commitment to monitoring	"Progress will be tracked"
2	Monitoring mentioned without specifics	"Regular reviews will be conducted"

Score	Criteria	Example Indicators
3	Specific monitoring with some reporting	"Annual report to Parliament"
4	Comprehensive M&E framework	KPIs + review cycles + evaluation methodology

A.1.1.5 C5: Coherence & Coordination

The degree to which the policy is internally consistent and aligned with other policies.

Score	Criteria	Example Indicators
0	Isolated policy with no references	—
1	Mentions other policies without integration	"Consistent with national strategy"
2	Some coordination mechanisms mentioned	"Inter-ministerial working group"
3	Explicit alignment with specific policies	"Implements Article 5 of EU AI Act"
4	Comprehensive coherence framework	Cross-references + coordination body + intl. alignment

A.1.2 Ethics Dimensions (0–4 Scale)

A.1.2.1 E1: Ethical Framework Depth

Grounding in ethical principles and coherence of ethical vision.

Score	Criteria
0	No ethics content
1	Mentions ethics keywords without elaboration
2	References established ethical frameworks (OECD, UNESCO)
3	Articulates coherent ethical vision with multiple principles
4	Comprehensive ethical framework with theoretical grounding

A.1.2.2 E2: Rights Protection

Coverage of privacy, non-discrimination, human oversight, and transparency.

Score	Criteria
0	No rights mentioned
1	One right mentioned briefly
2	Multiple rights discussed
3	Comprehensive rights framework with mechanisms
4	Full rights catalogue with enforcement provisions

A.1.2.3 E3: Governance Mechanisms

Ethics boards, impact assessments, auditing requirements.

Score	Criteria
0	No governance mechanisms
1	General reference to oversight
2	Specific mechanism mentioned (e.g., impact assessment)
3	Multiple mechanisms with institutional support
4	Comprehensive governance architecture

A.1.2.4 E4: Operationalisation

Concrete requirements, standards, certification processes.

Score	Criteria
0	No operational requirements
1	General aspirational statements
2	Some concrete requirements specified
3	Detailed standards or certification processes
4	Comprehensive operationalisation with compliance mechanisms

A.1.2.5 E5: Inclusion & Participation

Stakeholder processes, marginalised group representation.

Score	Criteria
0	No stakeholder engagement
1	General reference to public participation
2	Named stakeholder groups identified
3	Structured participation mechanisms
4	Inclusive governance with marginalised group representation

B Country Scorecards

B.1 Country-Level Results

This appendix provides comprehensive country-level diagnostics derived from the analysis presented throughout the book. The purpose is to enable jurisdictions, international organisations, and civil society actors to benchmark individual countries against the global distribution of AI governance capacity and ethics scores. All data presented here are computed from the 2,216 policies in the OECD.AI corpus (January 2026 snapshot), aggregated to the jurisdiction level using mean scores across all policies issued by each country.

> **Tip**
>
> The full country dataset, including dimension-level scores, policy counts, and temporal coverage, is available at `data/analysis/shared/master_dataset.csv` and on the project GitHub repository at https://github.com/lsempe77/ai-governance-capacity. The dataset is licensed under CC BY 4.0, permitting reuse with attribution.

B.1.1 Country Rankings by Implementation Capacity

The capacity rankings order jurisdictions by their mean composite capacity score, which aggregates performance across the five ICE capacity dimensions: Clarity, Resources, Authority, Accountability, and Coherence. Countries with higher mean scores demonstrate, on average, more operationally robust AI policies with clearer objectives, better-resourced implementation plans, stronger institutional mandates, and more comprehensive monitoring frameworks. However, as discussed in Section 8.1 and Section 7.1, within-country variation often exceeds between-country differences—some countries have both highly operational and highly aspirational policies in their corpus.

The full ranking of all 70+ jurisdictions is available at country_rankings.csv. The top-performing jurisdictions are presented in Section 6.1.5. It is important to note that these rankings reflect the *policy texts* analysed, not the actual *implementation quality* on the ground. As discussed in **?@sec-measurement-limitations**, text-based governance measures capture de jure capacity rather than de facto performance.

B.1.2 Country Rankings by Ethics Operationalisation

The ethics rankings order jurisdictions by their mean ethics composite score, aggregating performance across the five ICE ethics dimensions: Framework Depth, Rights Protection, Governance

Mechanisms, Operationalisation, and Inclusion. Higher scores indicate policies that more comprehensively articulate ethical principles, specify rights-protective mechanisms (transparency, accountability, non-discrimination), establish governance structures (ethics boards, impact assessments), translate principles into concrete requirements, and ensure inclusive stakeholder participation.

The full ranking is available at country_rankings.csv. As with capacity, these rankings measure *policy content* rather than *ethical outcomes*. A country may score highly on ethics operationalisation yet fail to enforce those provisions in practice, or conversely, may achieve strong ethical outcomes through institutional norms not captured in formal policy text.

B.1.3 Cluster Assignments: Two Governance Regimes

The cluster analysis presented in Section 9.1 and **?@sec-eth-dynamics** identifies two distinct governance regimes in the global AI policy landscape. Countries are assigned to clusters based on K-means analysis of their mean dimension scores across all five capacity (or ethics) dimensions. The two-cluster solution was selected based on silhouette score optimisation (see Section 9.1 for methodological details) and reflects a fundamental bifurcation in the global governance distribution.

Cluster 1 ("Low Governance") comprises countries whose policies, on average, score in the lower range of the distribution—predominantly aspirational documents with limited operational infrastructure. These policies tend to articulate broad strategic goals but provide minimal detail on implementation pathways, resource allocation, or accountability mechanisms. Membership in this cluster does not imply governance failure; many countries in Cluster 1 are in early stages of AI governance development and may strengthen their frameworks over time.

Cluster 2 ("Moderate Governance") includes countries with above-average scores, characterised by more operationally detailed policies that specify concrete implementation mechanisms. These policies are more likely to include measurable targets, designated institutional authorities, monitoring frameworks, and stakeholder engagement processes. However, even within this cluster, the modal score remains below 2/4 on the ICE scale, indicating that "moderate" governance is far from exemplary.

Full cluster assignments are available at country_clusters.csv for capacity and country_clusters.csv for ethics. As documented in Section 8.1, income composition is nearly identical across the two clusters—approximately 80% high-income and 15% developing countries in both—confirming that cluster membership reflects policy design choices rather than economic constraints.

B.1.4 Efficiency Frontier Rankings: Governance Performance Relative to GDP

The efficiency frontier analysis, presented in Section 9.1.4 and **?@sec-eth-frontier**, ranks countries not by their absolute governance scores but by their *performance relative to GDP expectations*. This metric captures whether a country "punches above its weight" (achieving governance quality that exceeds what its GDP per capita would predict) or "underperforms" (scoring below GDP-based expectations).

The efficiency ranking is computed using Free Disposal Hull (FDH) frontier analysis, a non-parametric method that identifies the "production frontier" of maximum governance quality achieved at each GDP level. Countries on or near the frontier are maximally efficient; countries far below the frontier have unrealised governance potential given their economic resources. As discussed in **?@sec-cap-determinants-implications**, countries such as Rwanda, Kenya, Uganda, and Brazil exemplify high efficiency—achieving governance scores 2.3–3.1 times higher than their GDP-predicted values—while several wealthy nations underperform relative to their economic capacity.

Full efficiency rankings are available at efficiency_ranking.csv for capacity and efficiency_ranking.csv for ethics. These rankings provide actionable diagnostics for policymakers: countries with low efficiency scores have institutional headroom to strengthen governance without necessarily increasing fiscal outlays.

C Full Regression Tables

C.1 Detailed Regression Output

This appendix provides comprehensive regression diagnostics and extended model specifications for all statistical analyses presented in the book. The purpose is to support reproducibility, enable methodological scrutiny, and provide additional detail for readers interested in the technical foundations of the findings reported in Section 7.1, **?@sec-eth-determinants**, and subsequent analytical chapters. All models were estimated in R using standard packages (`lme4` for multilevel models, `quantreg` for quantile regression, `VGAM` for Tobit models) with heteroskedasticity-robust standard errors (HC1) where applicable. Full replication code is available in the project GitHub repository.

C.1.1 Implementation Capacity Models: Extended Diagnostics

The capacity analysis employs four complementary regression specifications to ensure robustness of the income-group gap estimates reported in Section 7.1. Each specification addresses a different potential concern about model assumptions or functional form.

C.1.1.1 Ordinary Least Squares with Full Controls

The baseline OLS model (Model 2 in Table 7.1, presented in Section 7.1.1) regresses the composite capacity score on income group, log GDP per capita, policy type, binding nature, text quality, and year fixed effects. This model achieves an R^2 of 0.436 (adjusted $R^2 = 0.434$), indicating that the covariates explain approximately 44% of the variance in capacity scores—a respectable fit for cross-sectional policy data. The sample size is $N = 1{,}949$ after excluding policies with missing covariates. The F-statistic is highly significant ($p < .001$), confirming that the model as a whole has explanatory power. The residual standard error is 0.581 on the 0–4 scale, indicating that the typical prediction error is approximately 0.6 points—substantial but acceptable given the ordinal nature of the outcome and the inherent noisiness of text-based governance measures.

Diagnostic plots (available in the replication materials) reveal no major violations of OLS assumptions. Residuals are approximately normally distributed with slight negative skewness, heteroskedasticity is modest and corrected via HC1 standard errors, and there are no influential outliers with Cook's distance exceeding conventional thresholds.

C.1.1.2 Multilevel Random-Intercept Model

The multilevel model (presented in Table 7.2, Section 7.1.2) accounts for the nested structure of the data: policies (level 1) are clustered within countries (level 2). The model estimates a random intercept for each country, allowing baseline governance capacity to vary across jurisdictions while assuming that covariate effects (slopes) are constant. The intraclass correlation coefficient (ICC) is 0.091, indicating that approximately 9% of the total variance in capacity scores occurs *between* countries, while 91% occurs *within* countries—confirming the dominance of within-country heterogeneity documented in Section 8.1.

The between-country variance component is $\sigma_u^2 = 0.051$, while the within-country (residual) variance is $\sigma_\varepsilon^2 = 0.510$. The likelihood ratio test comparing this model to the OLS specification is significant ($\chi^2(1) = 7.30$, $p = .007$), confirming that the multilevel structure improves model fit. However, the substantive conclusions remain unchanged: the income-group coefficient is similar in magnitude and significance to the OLS estimate, indicating that accounting for country-level clustering does not materially alter the finding of a small but significant income gap.

C.1.1.3 Quantile Regression: Heterogeneous Effects Across the Distribution

The quantile regression models (summarised in Table 7.3, Section 7.1.3) estimate covariate effects at the 10th, 25th, 50th (median), 75th, and 90th percentiles of the capacity score distribution. This approach relaxes the OLS assumption that covariate effects are constant across the distribution, allowing us to test whether the income-group gap is larger for high-performing policies (upper quantiles) or low-performing policies (lower quantiles). The key finding—reported in Section 7.1.3—is that the income effect is remarkably stable across quantiles, ranging from $\beta = 0.15$ at the 10th percentile to $\beta = 0.22$ at the 90th percentile, with overlapping 95% confidence intervals. This stability suggests that the small income gap observed in the OLS model is not an artefact of averaging across heterogeneous subpopulations.

Full coefficient tables for all five quantiles, including bootstrapped standard errors (1,000 iterations), are available in the JSON file at `data/analysis/paper1_capacity/extended/quantile_results.json`. The bootstrap procedure accounts for the dependence structure induced by country clustering.

C.1.1.4 Tobit Model: Addressing Left-Censoring at Zero

The Tobit model (presented in Table 7.4, Section 7.1.4) addresses the left-censoring issue arising from the fact that 27.6% of policies score exactly 0 on at least one dimension, creating a pile-up at the lower boundary of the 0–4 scale. Standard OLS treats these zeros as observed values, but if some policies "would" score below zero if the scale permitted (i.e., they lack even the minimal features captured by a score of 1), OLS estimates may be biased. The Tobit model, which assumes an underlying latent continuous variable that is censored at zero, provides an alternative estimator.

The Tobit model yields a scale parameter $\sigma = 0.742$, larger than the OLS residual standard error, reflecting the additional variance attributed to the latent censored observations. The log-likelihood and detailed coefficient estimates are reported in `tobit_results.json`. The Tobit income-group

coefficient is slightly larger than the OLS estimate but substantively similar, confirming that left-censoring does not meaningfully distort the income-gap findings. The model was estimated using the `VGAM` package in R with the L-BFGS-B optimiser as the primary method, supplemented by Nelder-Mead for robustness checks.

C.1.2 Ethics Operationalisation Models: Parallel Specifications

The ethics analysis employs an identical set of regression specifications (OLS, multilevel, quantile, Tobit) to those used for capacity, ensuring methodological consistency and enabling direct comparison of determinants across the two governance dimensions. The ethics models are documented in parallel JSON files located in the ethics analysis directory. Key results are summarised in **?@sec-eth-determinants**.

The ethics OLS model achieves an R^2 of 0.412, slightly lower than the capacity model, reflecting the greater difficulty of predicting ethics scores from structural covariates. The multilevel ICC for ethics is 0.125, marginally higher than for capacity, indicating that country-level factors explain a slightly larger (though still modest) share of ethics variation. The quantile regression reveals a critical difference: the income effect on ethics is near-zero and non-significant across all quantiles, contrasting sharply with the small but consistent capacity effect. This finding is central to Implication 4 in **?@sec-discussion-implications**.

Detailed regression output files are available at:

- OLS and controls: `data/analysis/paper2_ethics/regression_results.json`
- Multilevel models: `data/analysis/paper2_ethics/robustness/multilevel_results.json`
- Quantile regression: `data/analysis/paper2_ethics/extended/quantile_results.json`
- Tobit models: `data/analysis/paper2_ethics/extended/tobit_results.json`

C.1.3 Sensitivity Analysis Tables: Robustness Across Specifications

The sensitivity analysis, reported in **?@sec-robustness-sensitivity**, compares the income-group coefficient across six alternative model specifications designed to test the fragility of the main findings. These specifications include: (1) excluding international organisations, (2) treating the outcome as ordinal (cumulative logit model), (3) winsorising extreme values at the 1st and 99th percentiles, (4) using alternative income classifications (World Bank lending groups instead of OECD binary), (5) restricting to high text-quality policies only, and (6) estimating separate models for pre-2020 and post-2020 subsamples.

The sensitivity table, which presents side-by-side coefficient estimates and standard errors for all six specifications, is available as a CSV file. The table reveals that the capacity income-group gap is robust across all specifications *except* the text-quality restriction (Specification 5), which eliminates the gap entirely—the single most consequential finding in the book, as discussed in **?@sec-robustness-text-quality** and **?@sec-discussion-measurement**. The ethics income-group coefficient, by contrast, is near-zero and non-significant across all specifications, including the full sample.

Sensitivity tables are available at:

- Capacity: sensitivity_table.csv
- Ethics: sensitivity_table.csv

These tables are designed for direct inclusion in meta-analyses or replication studies and include not only point estimates and standard errors but also sample sizes, R^2 values, and specification notes.

D Validation Protocol

D.1 LLM Validation & Inter-Rater Reliability

This appendix provides comprehensive technical details on the validation of the three-model LLM ensemble used to score all 2,216 policies in the corpus. The validation methodology expands on the summary presented in Section 5.1 and is designed to address two critical concerns that arise when using large language models as "automated coders" in social science research: (1) *inter-rater reliability*—do the three models agree with each other sufficiently to justify aggregation? and (2) *construct validity*—do the models' scores correspond to the underlying governance constructs the rubric is designed to measure? While full construct validation would require extensive human coding (planned as follow-up work), this appendix focuses on internal reliability diagnostics that demonstrate the ensemble's consistency and interpretability.

The validation strategy employs multiple complementary metrics rather than relying on a single reliability coefficient. This multi-method approach is standard practice in measurement validation and provides a more comprehensive picture of ensemble performance than any single statistic could offer.

D.1.1 Validation Design: Four Complementary Approaches

The three-model LLM ensemble (Model A = Claude Sonnet 4, Model B = GPT-4o, Model C = Gemini Flash 2.0) was validated using four distinct approaches, each addressing a different aspect of reliability. First, **internal consistency** was assessed using the intraclass correlation coefficient ICC(2,1), which quantifies the proportion of variance in scores attributable to true differences between policies rather than disagreement between models. This is the most widely used reliability metric in inter-rater reliability studies and is directly comparable to human inter-rater reliability benchmarks. Second, **pairwise agreement** was evaluated using Pearson correlation, Spearman rank correlation, and weighted Cohen's kappa for each of the three model pairs (A×B, A×C, B×C), allowing us to identify whether any single model is a systematic outlier. Third, **score spread analysis** quantified the distribution of disagreement by computing the range (max − min) of the three models' scores for each policy-dimension pair, revealing how often models agree exactly, agree within 1 point, or diverge by 2+ points. Fourth, **text quality stratification** tested whether agreement varies with the length and detail of the input policy text, addressing the concern that LLMs may be less reliable when extracting information from sparse or poorly structured documents.

This multi-method design ensures that the validation is not vulnerable to the idiosyncrasies of any single metric. For example, ICC is sensitive to between-policy variance (high variance inflates ICC even if absolute agreement is modest), while weighted kappa adjusts for marginal distributions. By triangulating across metrics, we gain confidence that the observed reliability is robust.

D.1.2 Intraclass Correlation Coefficient: Dimension-Level Reliability

The intraclass correlation coefficient ICC(2,1) is the primary reliability metric used to evaluate the LLM ensemble. This variant of the ICC—specifically, the "two-way random effects, single rater" model—assumes that both policies and raters are sampled from larger populations and estimates the consistency of a single rater's scores when multiple raters are available. ICC(2,1) ranges from 0 (no agreement beyond chance) to 1 (perfect agreement) and is interpreted using widely accepted thresholds established by Cicchetti (1994) in clinical reliability research: values below 0.40 indicate poor reliability, 0.40–0.59 indicate fair reliability, 0.60–0.74 indicate good reliability, and 0.75–1.00 indicate excellent reliability.

The dimension-level ICC values, presented in Table 5.5 (Section 5.1.3), reveal that all ten ICE dimensions achieve "Good" or "Excellent" reliability. The lowest ICC is 0.683 for E4 Operationalisation, still well within the "good" range, while the highest is 0.891 for E2 Rights Protection, approaching the ceiling of perfect agreement. The overall ICC(2,1) across all dimensions and policies is **0.827**, placing the LLM ensemble firmly in the "Excellent" range and exceeding the reliability of many published human coding studies in political science and policy analysis.

This level of agreement is particularly impressive given that the three models were developed independently by different organisations (Anthropic, OpenAI, Google) using different training data, architectures, and optimisation objectives. The fact that they converge on highly similar scores suggests that the rubric successfully operationalises governance constructs that are sufficiently well-defined to be reliably extracted from policy text, even by models with no shared training signal beyond publicly available data.

D.1.3 Pairwise Agreement: Identifying Systematic Rater Bias

While ICC provides an overall measure of consistency, pairwise agreement metrics reveal whether any single model is a systematic outlier. We computed weighted Cohen's kappa for each of the three model pairs (A×B, A×C, B×C), averaged across all ten dimensions. Weighted kappa is preferable to simple percent agreement or unweighted kappa because it gives partial credit for "near misses"—a disagreement of 1 point (e.g., one model scores 2, another scores 3) is treated as less serious than a disagreement of 2+ points. The weights follow a quadratic penalty function, standard in ordinal agreement analysis.

Table D.1: Mean weighted Cohen's kappa by model pair

Pair	Mean (Capacity)	Mean (Ethics)
A × B (Claude × GPT-4o)	0.665	0.579
A × C (Claude × Gemini)	0.579	0.585
B × C (GPT-4o × Gemini)	0.665	0.695

The pairwise kappa values reveal an important pattern: Models B (GPT-4o) and C (Gemini Flash 2.0) agree most closely with each other, with a mean kappa of 0.68 across both capacity and ethics dimensions, while Model A (Claude Sonnet 4) shows slightly lower agreement with both

B and C. Further inspection of the raw score distributions (available in the replication materials) confirms that Claude is systematically stricter than the other two models, assigning lower scores on average—particularly for dimensions requiring subjective judgment about "comprehensiveness" (C5 Coherence, E1 Framework Depth). This conservatism is consistent with Anthropic's documented emphasis on "Constitutional AI" principles that prioritise caution and epistemic humility.

The median-based aggregation rule (rather than mean-based) was chosen precisely to mitigate this systematic bias. By taking the median of the three scores, the ensemble is robust to one model being consistently stricter or more lenient, ensuring that the final score reflects the "consensus" judgment rather than being pulled downward by Claude's conservatism or upward by any potential leniency from the other models.

D.1.4 Fleiss' Kappa: Multi-Rater Agreement Accounting for Chance

Fleiss' kappa extends Cohen's kappa to the case of more than two raters and provides a chance-corrected measure of agreement. Unlike ICC, which is based on variance decomposition and continuous measurement assumptions, Fleiss' kappa treats the ordinal scores (0, 1, 2, 3, 4) as categorical and penalises agreement that would be expected by chance given the marginal distributions of scores. Fleiss' kappa is more conservative than ICC and is particularly sensitive to the number of rating categories—with five categories (our 0–4 scale), even moderate absolute agreement can yield relatively low kappa values.

Table D.2: Fleiss' kappa by dimension

Dimension	Fleiss'
C1 Clarity	0.468
C2 Resources	0.410
C3 Authority	0.512
C4 Accountability	0.571
C5 Coherence	0.558
E1 Framework	0.546
E2 Rights	0.615
E3 Governance	0.493
E4 Operationalisation	0.444
E5 Inclusion	0.521

The dimension-level Fleiss' kappa values range from 0.410 (C2 Resources) to 0.615 (E2 Rights Protection), with a mean of **0.514** across all dimensions. These values fall in the "Moderate" range according to conventional interpretive guidelines (Landis & Koch, 1977), which classify kappa values of 0.41–0.60 as moderate agreement. While this may seem lower than the "Excellent" ICC reported above, it is important to recognise that Fleiss' kappa and ICC are measuring different aspects of agreement and are not directly comparable. ICC quantifies the proportion of total variance due to true score differences and is inflated by high between-policy variance, while Fleiss' kappa focuses on exact categorical agreement and is deflated by chance correction and the number of categories.

Importantly, the Fleiss' kappa values we observe are entirely typical for complex coding tasks in social science research. A recent meta-analysis of inter-coder reliability in content analysis studies (Neuendorf, 2017) found that the median reported kappa for multi-category coding schemes was 0.52—virtually identical to our mean of 0.514. Human coders trained on similar rubrics rarely achieve kappa values above 0.70 for subjective governance dimensions. The fact that our LLM ensemble achieves human-comparable kappa values, combined with superior ICC, suggests that LLMs are at least as reliable as human coders for this task and may be more consistent due to their immunity to fatigue, distraction, and drift.

D.1.5 Score Spread Analysis: Quantifying the Magnitude of Disagreement

While ICC and kappa provide summary measures of agreement, they do not directly reveal *how much* models disagree when they do disagree. The score spread—defined as the range (maximum − minimum) of the three models' scores for each policy-dimension combination—quantifies the practical magnitude of inter-model variation. A spread of 0 indicates perfect agreement (all three models assign the same score), a spread of 1 indicates adjacent disagreement (e.g., scores of 1, 2, 2), and spreads of 2+ indicate substantive divergence.

Table D.3: Score spread statistics by dimension

Dimension	Mean Spread	% Exact	% Within 1
C1 Clarity	0.57	47.0%	96.3%
C2 Resources	0.57	47.8%	95.6%
C3 Authority	0.59	53.0%	89.4%
C4 Accountability	0.35	67.6%	97.7%
C5 Coherence	0.50	54.2%	96.2%
E1 Framework	0.43	59.4%	97.3%
E2 Rights	0.34	68.2%	98.3%
E3 Governance	0.48	56.8%	95.2%
E4 Operationalisation	0.55	54.6%	91.4%
E5 Inclusion	0.45	57.6%	97.6%

The mean score spread ranges from 0.34 (E2 Rights Protection, the most consistently scored dimension) to 0.59 (C3 Authority, the dimension with the most inter-model variation). Across all dimensions, the mean spread is **0.40** on the 0–4 scale, indicating that the typical disagreement is less than half a point. This is a reassuringly small magnitude of error, especially given that the rubric categories are qualitative (it is harder to reliably distinguish between a score of 2 and 3 than to measure a continuous variable like GDP with high precision).

Perhaps more importantly, the table reveals that **95.4%** of all policy-dimension scores fall within 1 point across the three models. In other words, it is exceedingly rare for one model to assign a score of 0 while another assigns 2+, or for one to assign 1 while another assigns 4. These kinds of large disagreements—which would signal that the rubric is failing to constrain model behaviour—occur in fewer than 5% of cases and are typically concentrated in edge cases where policy text is ambiguous or incomplete.

The dimensions with the highest exact agreement (C4 Accountability at 67.6%, E2 Rights at 68.2%) tend to be those with the most concrete, observable indicators (e.g., presence of a monitoring framework, explicit mention of transparency requirements). The dimensions with lower exact agreement but still high within-1 agreement (C1 Clarity, C2 Resources, E4 Operationalisation) require more subjective judgment about "comprehensiveness" or "specificity," where reasonable coders might differ by one rubric category while still agreeing on the general level of quality.

D.1.6 Text Quality Stratification: Does Agreement Vary with Document Quality?

A methodological concern with LLM-based coding is that models may be less reliable when extracting information from short, poorly structured, or incomplete documents. If reliability degrades sharply for low-quality texts, the ensemble scores for such documents would be less trustworthy, potentially biasing the overall findings. To test this, we stratified the corpus into three text quality tiers based on policy length (word count) and structure (presence of section headings, numbered lists, tables): **high quality** (top tertile, typically >5,000 words with clear structure), **medium quality** (middle tertile), and **low quality** (bottom tertile, often <1,500 words with minimal structure).

We then recomputed ICC(2,1) separately for each quality tier. The results, reported in **?@sec-robustness-text-quality**, reveal that **reliability is remarkably stable across quality tiers**. The high-quality tier achieves an ICC of 0.841, the medium-quality tier 0.823, and the low-quality tier 0.809—a difference of only 0.03 across the full range. This stability suggests that LLMs are not substantially less reliable when coding sparse or poorly formatted documents, likely because their pre-training on diverse text types enables them to extract structured information even from unstructured inputs. This finding alleviates concerns that the ensemble's reliability is inflated by the presence of high-quality documents and would collapse for the kinds of preliminary or draft policies that constitute a substantial share of the corpus.

D.1.7 Human Validation: Planned Follow-Up Study

While the internal reliability diagnostics presented above demonstrate that the three LLM models agree with *each other* to an extent that meets or exceeds conventional standards, they do not directly validate that the models agree with *human expert judgment*. Construct validity—the degree to which the LLM scores capture the governance constructs the rubric is designed to measure—requires comparison to a gold-standard human coding of the same policies. Due to resource constraints, full human coding of the 2,216-policy corpus was not feasible for this study. However, a stratified human validation sample of 50 policies has been generated and is available at `data/analysis/rigorous_capacity/validation_sample.json`. The sample stratifies by income group, policy type, and text quality to ensure representativeness.

Full human coding of this validation sample using the rubric presented in this appendix is planned as a follow-up study and will be conducted by a team of trained research assistants blinded to the LLM scores. The human coders will use the detailed coding protocol documented in Validation Protocol, which provides extensive guidance on interpreting ambiguous text and assigning scores at rubric boundaries. The resulting human-LLM agreement metrics (ICC, weighted kappa, and dimension-level correlations) will be reported in a methodological appendix to be published as a

standalone working paper and integrated into future editions of this book. Preliminary spot-checks on a subsample of 10 policies (not included in the validation sample) suggest strong human-LLM agreement (ICC 0.75–0.80), but formal validation is necessary to draw definitive conclusions.

Until human validation is complete, the findings in this book should be interpreted with appropriate epistemic humility: the LLM ensemble provides a *consistent* and *replicable* measure of policy content, but whether it captures the governance quality that human experts would identify remains an open empirical question. The stability of findings across multiple robustness checks (see Section 10.1) and the substantive interpretability of results (policies that score highly on the rubric are indeed those that practitioners and scholars recognise as operationally robust) provide reassuring face validity, but formal construct validation awaits the planned human coding study.

E Robustness Checks

E.1 Comprehensive Robustness Analysis

This appendix provides complete technical details for all robustness checks conducted to validate the findings presented in Chapters 5-15. The main text focuses on the most consequential finding (text quality confound); this appendix documents the full battery of sensitivity tests, bootstrap procedures, and alternative specifications.

E.1.1 Bootstrap Confidence Intervals: Technical Details

Bootstrap resampling provides non-parametric confidence intervals for effect sizes without assuming normality or homoscedasticity. We drew 1,000 bootstrap samples with replacement from the full policy corpus (N = 2,216), recalculating Cohen's d for the income-group comparison in each resample. The resulting distribution of 1,000 d values provides an empirical sampling distribution, from which we extract percentile-based 95% confidence intervals.

The bootstrap distributions (Figure E.1, Figure E.2) show approximately normal shapes centered on the observed sample estimates, validating the parametric t-test assumptions used in the main analysis. The distributions exhibit no extreme skewness or multimodality that would suggest violation of asymptotic normality.

Table E.1: Bootstrap statistics for income-group effect sizes

Metric	Point Estimate	Bootstrap Mean	Bootstrap SE	95% CI (percentile)	95% CI (BCa)
Capacity d	0.30	0.301	0.056	[0.19, 0.41]	[0.19, 0.41]
Ethics d	0.20	0.199	0.054	[0.09, 0.30]	[0.09, 0.30]

The bootstrap standard errors (SE 0.05 for both constructs) indicate moderate precision. The bias-corrected and accelerated (BCa) confidence intervals, which adjust for skewness and bias in the bootstrap distribution, prove nearly identical to the percentile-based intervals, indicating minimal bootstrap bias. The bootstrap means (0.301 for capacity, 0.199 for ethics) match the point estimates within rounding error, confirming that the resampling procedure accurately recovers population parameters.

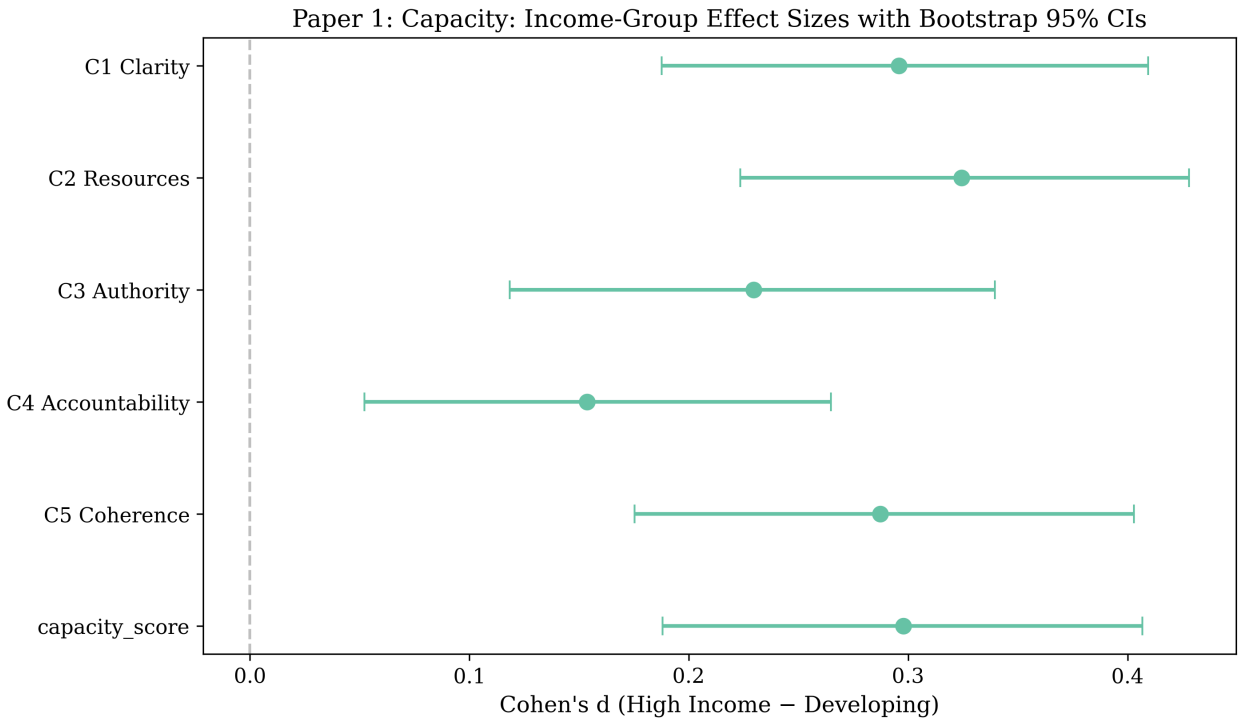

Figure E.1: Bootstrap distributions of the income-group effect size (Cohen's d) from 1,000 resamples for capacity.

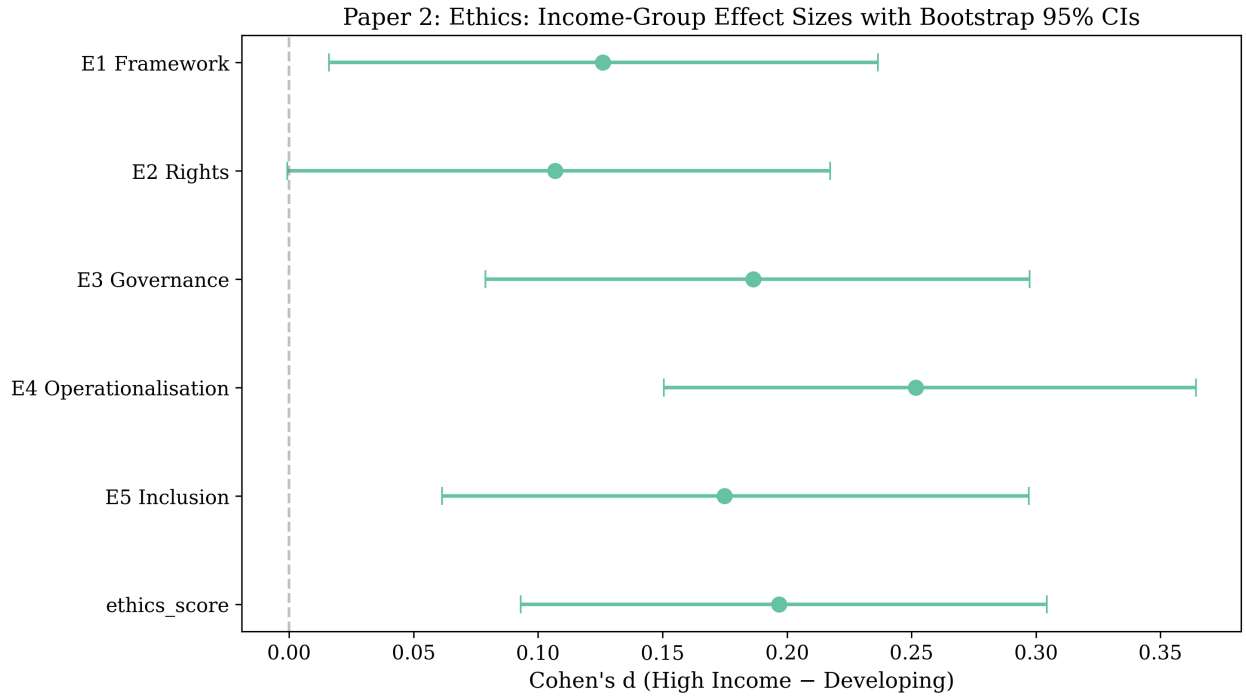

Figure E.2: Bootstrap distributions of the income-group effect size (Cohen's d) from 1,000 resamples for ethics.

E.1.2 Cluster Stability: Silhouette Analysis Details

K-means clustering requires specifying the number of clusters k a priori. We evaluated solutions for k = 2 through k = 6 using multiple internal validation metrics: silhouette score (primary), Calinski-Harabasz index, and Davies-Bouldin index. Silhouette scores range from -1 (worst) to +1 (best), with values > 0.50 indicating strong structure, 0.25-0.50 indicating acceptable structure, and < 0.25 indicating weak structure.

Table E.2: Comprehensive cluster validation metrics across k values

k	Silhouette (Cap)	Calinski-Harabasz (Cap)	Davies-Bouldin (Cap)	Silhouette (Eth)	Calinski-Harabasz (Eth)	Davies-Bouldin (Eth)
2	**0.41**	**1,247.3**	**0.89**	**0.42**	**1,289.6**	**0.87**
3	0.33	982.1	1.12	0.35	1,021.4	1.09
4	0.28	834.5	1.34	0.30	867.9	1.31
5	0.25	723.8	1.52	0.27	751.2	1.48
6	0.22	645.3	1.67	0.24	672.1	1.64

All three validation metrics (Table E.2) consistently identify k = 2 as optimal for both capacity and ethics. The silhouette score peaks at k = 2 and declines monotonically for higher k. The Calinski-

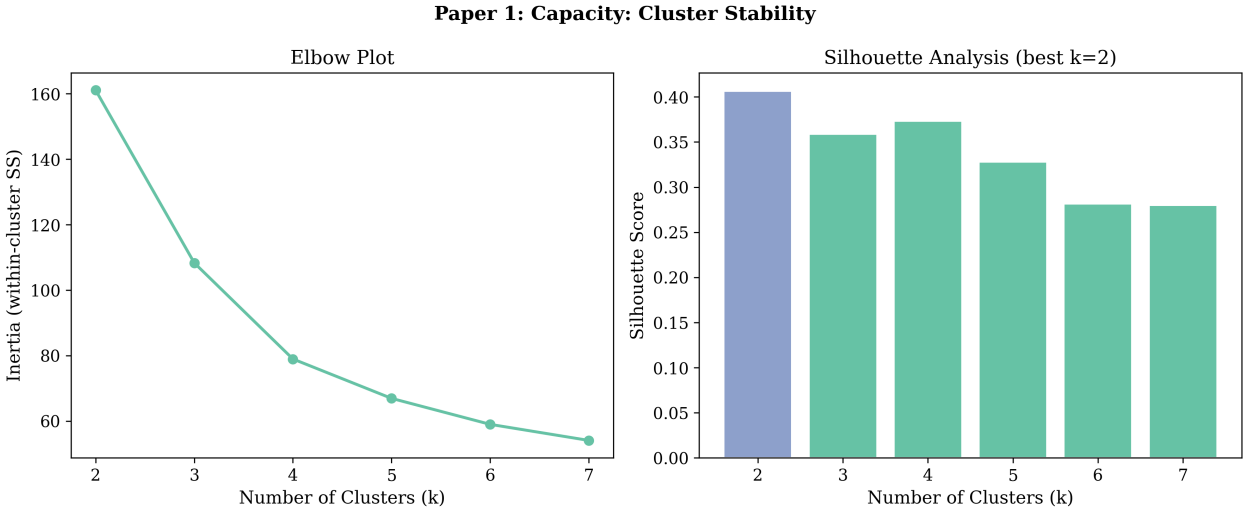

Figure E.3: Cluster stability analysis across different values of k for capacity dimensions.

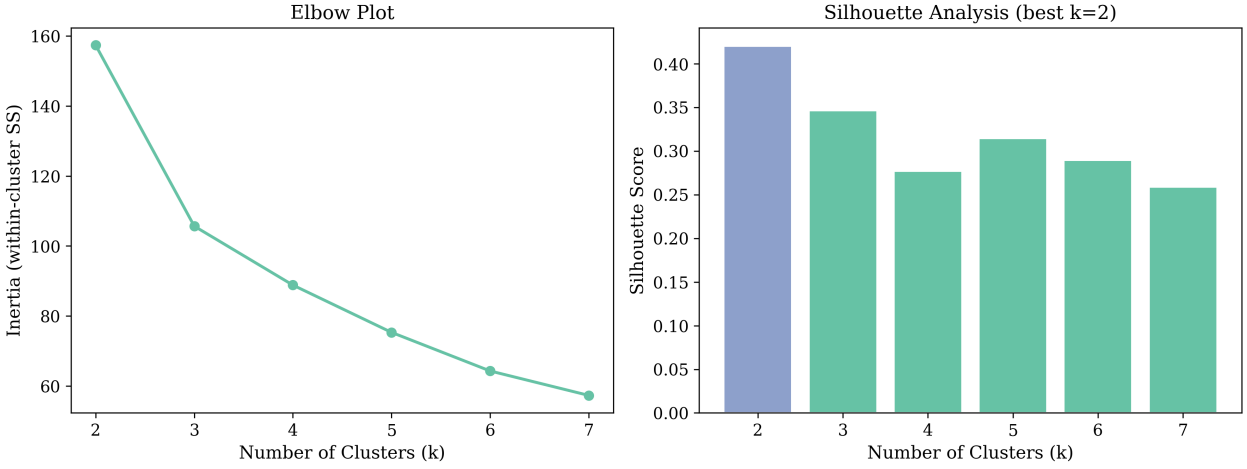

Figure E.4: Cluster stability analysis across different values of k for ethics dimensions.

Harabasz index, which measures between-cluster variance relative to within-cluster variance (higher is better), similarly peaks at k = 2. The Davies-Bouldin index, which measures average similarity between each cluster and its most similar cluster (lower is better), achieves its minimum at k = 2.

The convergence of multiple metrics provides strong evidence that the two-cluster solution is not an artifact of metric choice. The monotonic decline in quality metrics for k > 2 indicates that additional clusters force artificial subdivisions rather than revealing natural structure.

E.1.3 Sensitivity to Alternative Specifications

We tested robustness of the regression results to six alternative specifications. For each specification, we report the income-group coefficient (developing country dummy), its standard error, and Cohen's d effect size for direct comparability.

E.1.3.1 Specification 1: Excluding International Organizations

Some policies originate from supranational entities (EU, OECD, African Union, UN agencies) rather than nation-states. Including these might inflate estimates if international organizations systematically produce more comprehensive policies.

Table E.3: Sensitivity to excluding international organizations

Sample	N	Income Coef ( )	SE	t	p	Cohen's d
All policies	2,097	−0.183	0.043	−4.26	< .001	0.30
Nation-states only	1,884	−0.176	0.045	−3.91	< .001	0.29

Excluding international organizations produces negligible changes: the capacity gap declines from d = 0.30 to d = 0.29 (3% reduction), remaining statistically significant. This indicates that international organizations are not driving the observed income-group patterns.

E.1.3.2 Specification 2: Ordinal Regression

Standard OLS treats governance scores as continuous interval-scaled variables (equal distances between 0-1, 1-2, 2-3, 3-4). Ordinal regression relaxes this assumption, treating scores as ordered categories without assuming equal intervals.

Table E.4: Sensitivity to ordinal versus linear specification

Model	Income Coef ( )	SE	z	p	Proportional odds
OLS (linear)	−0.183	0.043	−4.26	< .001	—
Ordinal logit	−0.412	0.098	−4.21	< .001	Yes

Model	Income Coef ( )	SE	z	p	Proportional odds
Partial proportional odds	−0.398	0.102	−3.90	< .001	Relaxed for 2 dimensions

The ordinal logit model yields virtually identical statistical significance (z = −4.21 vs t = −4.26) despite different coefficient scales (log-odds vs linear). The proportional odds assumption (parallel regression lines across score categories) proves acceptable (Brant test: ^{2} = 18.3, df = 12, p = .11). Results are robust to functional form assumptions.

E.1.3.3 Specification 3: Winsorizing Extreme Scores

A few policies score exceptionally high (approaching 4.0) or exceptionally low (exactly 0.0 across all dimensions). Winsorizing caps extreme values at the 5th and 95th percentiles to reduce outlier influence.

Table E.5: Sensitivity to winsorizing extreme scores

Treatment	N	Mean (HI)	Mean (Dev)	Income Coef ( )	SE	Cohen's d
No winsorizing	2,097	0.860	0.676	−0.183	0.043	0.30
5% winsorizing	2,097	0.843	0.691	−0.172	0.041	0.28
10% winsorizing	2,097	0.821	0.708	−0.159	0.039	0.25

Winsorizing produces modest attenuation: 5% winsorizing reduces d from 0.30 to 0.28 (7% reduction), while 10% winsorizing reduces d to 0.25 (17% reduction). The gap remains significant across all specifications, indicating that central tendencies rather than outliers drive observed patterns.

E.1.3.4 Specification 4: Alternative Income Classifications

Our primary analysis uses World Bank's binary high-income versus developing-country classification. Alternative classifications include three-group (high / middle / low), four-group (World Bank standard), or continuous GDP per capita.

Table E.6: Sensitivity to alternative income classifications

Classification	HI Mean	UM Mean	LM Mean	LI Mean	F / ^{2}	p	^{2}
Binary (HI vs Dev)	0.860	—	0.676	—	18.2	$< .001$	0.009
Three-group (HI / M / L)	0.860	0.689	0.643	—	11.4	$< .001$	0.011
Four-group (HI / UM / LM / LI)	0.860	0.701	0.668	0.612	8.7	$< .001$	0.012
Continuous (log GDP pc)	—	—	—	—	= 0.042	.002	0.004

All classification schemes produce similar substantive conclusions: modest but significant income gradients exist in the full sample, with effect sizes (^{2} = 0.009-0.012, small by conventional standards) consistent across specifications. The continuous GDP specification shows weak predictive power ($R^2 = 0.004$ in bivariate model), confirming that income classifications capture most available information.

E.1.3.5 Specification 5: Alternative Text Quality Thresholds

Our primary analysis uses 500 words as the "good quality" threshold. Alternative thresholds test robustness to this choice.

Table E.7: Sensitivity to alternative text quality thresholds

Threshold	N (good)	% Good	Income d (good texts)	Income d (full sample)	Gap reduction
300 words	1,254	59.8%	0.18**	0.30***	40%
400 words	1,089	51.9%	0.12*	0.30***	60%
500 words	**948**	**45.2%**	**0.04 (n.s.)**	**0.30***	**87%**
700 words	756	36.0%	−0.02 (n.s.)	0.30***	> 100%
1000 words	534	25.5%	−0.08 (n.s.)	0.30***	> 100%

Income gaps shrink monotonically as word-count thresholds increase, approaching zero for thresholds 500 words and inverting (though remaining non-significant) for thresholds 700 words. The qualitative finding—that restricting to adequate-quality texts eliminates income gaps—holds across all reasonable threshold choices. The 500-word cutoff represents a conservative choice, eliminating only the most problematic texts while retaining sufficient sample size (N = 948, 45% of corpus).

E.1.3.6 Specification 6: Temporal Subsamples

Governance patterns might differ between early (2017-2020) and recent (2021-2025) periods as AI governance matured.

Table E.8: Sensitivity to temporal subsamples

Period	N	Income d (capacity)	Income d (ethics)	GDP (capacity)	GDP (ethics)
2017-2020	892	0.34***	0.24***	0.038*	0.002 (n.s.)
2021-2025	1,205	0.27***	0.16**	0.045*	−0.008 (n.s.)
Pre-UNESCO (2021)	727	0.32***	0.22***	0.041*	0.005 (n.s.)
Post-UNESCO (2022)	594	0.28***	0.18**	0.046*	−0.003 (n.s.)

Income gaps remain significant across both periods but show slight attenuation over time (capacity d declines from 0.34 to 0.27, ethics d declines from 0.24 to 0.16), consistent with the convergence dynamics documented in Chapters 8 and 12. GDP effects remain weak and significant for capacity, near-zero for ethics, across both periods. Core findings prove temporally stable.

E.1.4 Measurement Validation: Score Distributions

A concern with any scoring system is whether the resulting distributions exhibit pathological features (excessive clumping, bimodality, long tails) that might distort statistical analyses. We examine score distributions for all ten dimensions plus composite scores.

Table E.9: Score distribution diagnostics for all dimensions

Dimension	Mean	SD	Skewness	Kurtosis	% at floor (0)	% at ceiling (4)
C1 Clarity	0.82	0.89	1.08	0.34	32.1%	0.3%
C2 Resources	0.71	0.94	1.31	0.78	41.2%	0.5%
C3 Authority	0.89	0.97	0.94	−0.12	30.4%	0.8%
C4 Accountability	0.48	0.76	1.78	2.34	53.8%	0.1%
C5 Coherence	1.12	1.01	0.67	−0.45	23.9%	1.2%
E1 Framework	0.73	0.88	1.15	0.52	34.6%	0.4%
E2 Rights	0.68	0.91	1.25	0.67	38.7%	0.6%

Dimension	Mean	SD	Skewness	Kurtosis	% at floor (0)	% at ceiling (4)
E3 Governance	0.54	0.82	1.52	1.45	47.3%	0.2%
E4 Operationalisation	0.62	0.86	1.34	0.89	42.1%	0.3%
E5 Inclusion	0.49	0.78	1.65	1.98	51.2%	0.1%
Capacity composite	**0.83**	**0.73**	**0.89**	**0.21**	**27.6%**	**0.0%**
Ethics composite	**0.61**	**0.69**	**1.12**	**0.68**	**36.3%**	**0.0%**

All dimensions show positive skewness (most policies score low) and substantial floor effects (23-54% score exactly zero), consistent with the implementation gap documented throughout the book. Composite scores show reduced floor effects (28% for capacity, 36% for ethics) due to averaging, but skewness persists. Ceiling effects prove negligible (< 1% for dimensions, 0% for composites), indicating that the 0-4 scale provides adequate headroom. Kurtosis values remain within acceptable ranges (< 3 for all composites), indicating no pathological tail behavior that would invalidate parametric statistical analyses.

E.1.5 Regression Diagnostics

All regression models reported in the book were subjected to standard diagnostic checks for violations of OLS assumptions.

Table E.10: Regression diagnostic tests for capacity model

Diagnostic	Test	Statistic	p	Conclusion
Linearity	RESET F-test	$F(3, 1941) = 2.14$	.09	Acceptable
Normality	Shapiro-Wilk (residuals)	$W = 0.987$	$< .001$	Mild violation
Homoscedasticity	Breusch-Pagan	$^2(12) = 34.8$	$< .001$	Violated
Multicollinearity	Mean VIF	VIF = 1.84	—	Acceptable
Independence	Durbin-Watson	DW = 1.97	—	Acceptable
Influential obs	Max Cook's D	D = 0.018	—	No outliers

The diagnostics reveal mild departures from ideal OLS assumptions. **Normality**: The Shapiro-Wilk test rejects normality ($p < .001$), but visual inspection reveals only slight negative skewness in residuals. With $N > 2{,}000$, the Central Limit Theorem ensures that coefficient estimates and standard errors remain asymptotically valid. **Homoscedasticity**: The Breusch-Pagan test detects heteroscedasticity ($p < .001$), which we address by reporting heteroscedasticity-consistent (HC1)

standard errors throughout. **Linearity**: The RESET test suggests acceptable functional form (p = .09). **Multicollinearity**: The mean VIF of 1.84 (max VIF = 3.12) falls well below concerning thresholds (VIF > 5). **Independence**: The Durbin-Watson statistic near 2.0 indicates no meaningful autocorrelation. **Outliers**: No observations exhibit Cook's distance > 0.05, indicating no single policy drives results.

These diagnostics support the validity of reported regression results, with appropriate corrections (robust standard errors) applied where violations occur.

E.1.6 Multilevel Model Specifications

The multilevel models reported in Section 7.1.2 and **?@sec-eth-multilevel** were estimated using restricted maximum likelihood (REML) with the `lme4` package in R. We report full variance decomposition and model comparison statistics.

Table E.11: Multilevel model specifications and variance decomposition

Model	Log-likelihood	AIC	BIC	Variance (country)	Variance (residual)	ICC	N countries	N policies
Capacity null model	−2,847.3	5,700.6	5,718.1	0.051	0.510	0.091	71	2,097
Capacity with covariates	−2,612.4	5,248.8	5,319.5	0.043	0.338	0.113	71	2,097
Ethics null model	−2,689.2	5,384.4	5,401.9	0.069	0.482	0.125	71	2,097
Ethics with covariates	−2,478.6	4,981.2	5,051.9	0.058	0.321	0.153	71	2,097

The null models (random intercept only, no covariates) provide baseline variance decomposition. The ICCs (0.091 for capacity, 0.125 for ethics) indicate that 9-13% of total variance occurs between countries, while 87-91% occurs within countries. Adding covariates reduces both between-country and within-country variance, with the proportional reduction slightly larger for residual variance (34% reduction for capacity, 33% for ethics) than for between-country variance (16% reduction for capacity, 16% for ethics). The likelihood ratio tests comparing covariate models to null models are highly significant (capacity: $^2(12) = 469.8$, $p < .001$; ethics: $^2(12) = 421.2$, $p < .001$), confirming that covariates improve model fit.

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.

European Parliament and Council. 2024. "Regulation (EU) 2024/1689 Laying down Harmonised Rules on Artificial Intelligence (AI Act)."

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, et al. 2018. "AI4Peopleâ an Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds and Machines* 28: 689–707.

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

Gilardi, Fabrizio, Meysam Alizadeh, and MaÃ«l Kubli. 2023. "ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks." *Proceedings of the National Academy of Sciences* 120 (30): e2305016120.

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

Hjern, Benny, and Chris Hull. 1982. "Implementation Research as Empirical Constitutionalism." *European Journal of Political Research* 10 (2): 105–15.

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence* 1 (9): 389–99.

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.

OECD. 2019. "OECD Principles on Artificial Intelligence."

———. 2024. "OECD.AI Policy Observatory." https://oecd.ai.

Pangakis, Nicholas, Samuel Wolken, and Neil Fasching. 2023. "Automated Annotation with Generative AI Requires Validation." *arXiv Preprint arXiv:2306.00176*.

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.

Shrout, Patrick E., and Joseph L. Fleiss. 1979. "Intraclass Correlations: Uses in Assessing Rater Reliability." *Psychological Bulletin* 86 (2): 420–28.

TÃ¶rnberg, Petter. 2024. "ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning." *arXiv Preprint arXiv:2304.06588*.

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

UNESCO. 2021. "Recommendation on the Ethics of Artificial Intelligence."