AI Governance Observatory

Data, Methods, and Technical Appendices

Lucas Sempé

February 11, 2026

Table of contents

1 AI Governance Observatory

Data, Methods, and Technical Appendices

Preface

This volume documents the shared data infrastructure, scoring methodology, and technical validation underpinning three companion studies in the AI Governance Observatory:

1. **Book 1: AI Governance Implementation Capacity** — a cross-national analysis of whether AI policies contain the institutional infrastructure needed for execution
2. **Book 2: AI Ethics Governance Depth** — a cross-national analysis of how deeply policies operationalise ethical commitments
3. **Book 3: UNESCO AI Ethics Recommendation** — an assessment of how closely the global policy landscape aligns with UNESCO's 2021 framework

All three studies draw on the same corpus of **2,216 AI policies** from the OECD.AI Policy Observatory, scored by the same **three-model LLM ensemble** (Claude Sonnet 4, GPT-4o, Gemini Flash 2.0) using a shared 10-dimension rubric. Rather than repeat this shared methodology in each volume, we present it once here.

1.0.1 What this book contains

- **Data and Corpus Construction** (Section 2.1): How 2,216 policy entries were scraped, retrieved, and converted into 11.4 million words of analysis-ready text across 70+ jurisdictions
- **LLM Ensemble Scoring** (Section 3.1): The three-model scoring framework, inter-rater reliability (ICC = 0.827, "Excellent"), and composite score construction — including the UNESCO-specific alignment assessment
- **Technical Appendices**: The full scoring rubric (C1–C5, E1–E5), inter-rater validation protocol, comprehensive robustness checks, full regression tables, and country-level scorecards

1.0.2 What the companion books contain

Each results volume retains its own introduction, literature review, core analytical chapters, domain-specific robustness findings, discussion, and conclusion. Those volumes refer readers here for data and scoring details, keeping their focus on substantive findings.

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

2 Data and Methods

2.1 The OECD.AI Corpus

2.1.1 Corpus Construction

This study draws on the **OECD.AI Policy Observatory** (OECD 2024), the most comprehensive international tracker of AI policy initiatives. The Observatory catalogues government actions related to AI (national strategies, legislation, executive orders, guidelines, and programmes) with structured metadata on jurisdiction, year, policy type, target sectors, and responsible organisations. Each entry follows a consistent documentation schema, making it well suited for cross-national comparison.

The complete Observatory was scraped as of January 2026, obtaining **2,216 policy entries** across **70+ jurisdictions** spanning **2017–2025**.

Table 2.1: 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

The 70+ jurisdictions include not only major economies but also developing countries in Africa, Asia, and Latin America, providing the geographic diversity needed to examine governance gaps across income levels.

Document retrieval. The Observatory provides brief descriptions (typically <500 words) and links to source documents, but does not host full texts. To enable detailed assessment, a five-strategy cascading retrieval pipeline was constructed:

1. Direct download from `source_url` (~60% success)
2. Scraping embedded links from each OECD.AI entry page
3. Internet Archive Wayback Machine for moved/expired URLs
4. DuckDuckGo search with policy title, jurisdiction, and file type restrictions
5. Claude API web search for the most difficult cases

This pipeline achieved ~94% coverage (2,085 documents retrieved). The remaining entries, mostly press releases or brief announcements, were retained as OECD snippets.

Text extraction. Policy documents arrive in varied formats: text-based or scanned PDFs, web pages with complex navigation, documents ranging from single pages to hundred-page legislative texts. Format-specific extraction tools were employed: PyMuPDF (`fitz`) for PDFs, `trafilatura` for HTML content extraction, and OECD snippet text as a fallback.

Each document was classified into three quality tiers by word count:

Table 2.2: 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

About 80% of the corpus (1,754 documents) has enough text for reliable analysis. The 462 stubs contribute little analytically but are retained in corpus statistics. Total extracted text: 11.4 million words; median document length: 1,247 words (IQR: 318–4,892).

2.1.2 Sample and Metadata

The pipeline produces a unified corpus file (`corpus_enriched.json`) merging OECD metadata with extracted text and quality assessments. Each of the 2,216 entries retains its original OECD metadata (title, jurisdiction, year, URL, policy type, target sectors) plus extracted full text, quality classification, word count, and extraction method.

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, a binary classification contrasts 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). GDP per capita (current US dollars, 2023) serves as a continuous measure of economic development.

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.

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

Table 2.3: 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+**

The compositional imbalance is clear: high-income countries account for 77% of policies, developing countries for 18%. This reflects the actual distribution of AI governance activity. Rich countries have simply produced more policies and maintained more accessible archives. The robustness appendix (Section C.1) addresses this imbalance through sensitivity analyses including restriction to well-documented policies and country-level aggregations, and each companion volume reports domain-specific robustness checks.

2.1.3 Analytical Pipeline Overview

Figure 2.1 shows the full journey from raw OECD.AI metadata to the analytical outputs in the companion volumes. The 6,641 LLM API calls represent three model assessments for each of the 2,216 policies across 10 dimensions.

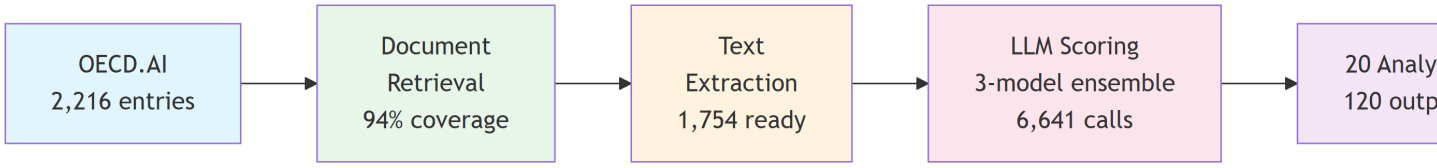

Figure 2.1: Analytical pipeline from corpus to results

2.1.4 Analytical Methods

The companion volumes use several complementary statistical methods. The core techniques are overviewed here; model specifications appear in the relevant chapters.

Text-to-data conversion. The core method uses frontier LLMs as policy analysts rather than relying on keyword extraction or topic models. Each model reads the full policy document, applies the scoring rubric across all 10 dimensions, and returns structured JSON scores with textual evidence. The three-model ensemble (Claude Sonnet 4, GPT-4o, Gemini Flash 2.0) uses the median score as the final assessment. ICC(2,1) = 0.827 indicates excellent inter-rater reliability; details are in Section 3.1.

Descriptive analysis. Each analytical section begins with descriptive statistics and visual exploration: dimension-specific histograms, ridge plots across groups, radar charts, box-and-violin overlays, and heatmaps.

Regression models. Chapters examining determinants of governance quality employ four complementary regression approaches: standard OLS to establish baseline relationships; multilevel models with random intercepts for countries to account for nested structure; quantile regression to examine heterogeneous effects across the distribution; and Tobit models to address the substantial floor effect (27.6% of policies score exactly zero) through left-censoring at zero.

Inequality analysis. Gini coefficients and Lorenz curves quantify overall inequality in governance scores. Theil's T index enables exact additive decomposition of total inequality into between-group (high-income vs. developing) and within-group components. Policy portfolio analysis examines breadth (whether countries address all dimensions) versus depth (score levels within covered dimensions).

Temporal analysis. Panel data methods separate within-country trends from between-country differences. First-difference models remove country fixed effects; Cohen's d effect sizes assess substantive significance; convergence analysis tests whether income-group gaps are narrowing, widening, or stable.

Multivariate methods. PCA examines the latent structure underlying the 10 governance dimensions, testing whether capacity and ethics represent empirically distinct constructs. Cronbach's alpha assesses internal consistency. K-means clustering identifies natural policy groupings, with optimal k determined through silhouette coefficients and bootstrap stability analysis.

Hypothesis testing. Welch's t-tests and Mann-Whitney U tests for group comparisons, chi-square tests for categorical associations, with exact p-values, effect sizes (Cohen's d, Cramér's V), and confidence intervals throughout.

2.1.5 Limitations and Reproducibility

The OECD.AI Observatory, while the most comprehensive international tracker available, introduces several systematic biases that readers should bear in mind throughout the companion volumes.

English-language dominance. The Observatory's documentation practices favour English-language sources. Policies originally published in English tend to receive fuller descriptions, more accessible source links, and more detailed metadata. Policies from Francophone Africa, the Arab states, or East Asian countries may appear as brief summaries even when extensive original-language documents exist. The LLM scoring models, while multilingual, perform best on English text. This creates a measurement pathway from language of publication → text quality → governance score that may systematically disadvantage non-Anglophone jurisdictions.

OECD member-state reporting bias. OECD member countries have institutional incentives and administrative capacity to report their policy activities to the Observatory. Non-member countries—particularly low-income countries with limited international engagement—may have governance instruments that simply do not appear in the database. The 77% high-income composition of the corpus (Table 2.3) likely reflects this reporting asymmetry as much as it reflects the actual distribution of AI governance activity.

Sub-national exclusion. The Observatory focuses on national-level policies, largely excluding state, provincial, and municipal AI governance. In federal systems—the United States, India, Brazil, Germany, Nigeria—substantial governance activity occurs at sub-national levels. California's AI regulation, for example, rivals many national frameworks in sophistication but does not appear in the corpus. This exclusion systematically understates governance capacity in federal countries and may distort cross-national comparisons.

Temporal coverage unevenness. Earlier years (2017–2019) contain fewer entries, particularly from developing countries that adopted AI governance later. Temporal analyses should be interpreted cautiously for this period, as apparent trends may reflect expanding Observatory coverage rather than genuine policy dynamics.

Classification quality. The Observatory's own categorisation of policy types, target sectors, and responsible organisations introduces noise. Some entries conflate national strategies with implementation programmes; others classify regional frameworks as national policies. These classification decisions, made by OECD analysts rather than by this study, propagate through all subsequent analyses.

UNESCO-specific limitation. The Observatory was not designed to track UNESCO alignment specifically. Policies that explicitly reference and implement the UNESCO Recommendation may not be tagged differently from policies developed independently. The alignment measurement in Book 3 thus captures *substantive overlap* with UNESCO components rather than *intentional implementation* of the Recommendation.

These limitations do not invalidate the analysis—the OECD.AI Observatory remains the best available data source for cross-national AI governance comparison—but they counsel caution when interpreting apparent gaps between countries, particularly between high-income and developing nations. The robustness appendix (Section C.1) and each companion volume's robustness chapter address the most consequential of these biases directly.

Reproducibility. All code is available at https://github.com/lsempe77/ai-governance-capacity. The pipeline uses deterministic document IDs (`MD5(url)[:12]`) for reproducibility. API calls used fixed model identifiers and structured JSON output schemas.

Use of Large Language Models. LLMs play two roles in this project, and I want to be upfront about both.

As the analytical instrument: Claude Sonnet 4, GPT-4o, and Gemini Flash 2.0 are the core measurement tool. They convert policy documents into structured scores. This is the methodology itself, documented in Section 2.1 and Section 3.1, with all scores preserved in the public repository.

As a writing aid: I used GitHub Copilot and Claude for drafting and editing assistance during manuscript preparation. I reviewed and revised all LLM-generated text; analytical decisions, interpretations, and intellectual contributions are my own.

3 LLM Ensemble Scoring and Validation

3.1 Measuring Governance Quality at Scale

3.1.1 Scoring Framework

Converting 2,216 policy documents into analyzable data required a framework capable of evaluating governance quality across diverse policy types, jurisdictions, and governance traditions. This study employs a **10-dimension assessment** organized into two domains: five capacity dimensions (implementation feasibility) and five ethics dimensions (ethical commitment operationalization). A separate **UNESCO alignment assessment** scores each policy against 25 components drawn from the 2021 Recommendation on the Ethics of AI.

Each dimension is scored 0–4, where 0 means the feature is absent and 4 indicates comprehensive operationalization with concrete mechanisms. The five-point scale balances granularity against reliability. Finer scales would add noise; coarser ones would obscure real differences.

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

Table 3.1: Capacity scoring dimensions

Code	Dimension	What It Measures
C1	Clarity & Specificity	Clear objectives, measurable targets, defined scope
C2	Resources & Budget	Dedicated funding, staffing, infrastructure
C3	Authority & Enforcement	Legal mandate, penalties, compliance mechanisms
C4	Accountability & M&E	Reporting, evaluation, oversight bodies
C5	Coherence & Coordination	Cross-agency alignment, international coordination

The mapping is intentional: Clarity tracks Mazmanian and Sabatier's clear-objectives condition, Resources captures Grindle's fiscal and technical requirements, Authority reflects legal structuring, Accountability addresses Lipsky's concern with constraining discretion, and Coherence captures the coordination challenges identified by Hjern and Hull (1982). These dimensions are analysed in depth in Book 1 (AI Governance Implementation Capacity).

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 3.2: 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 from Jobin, Ienca, and Vayena (2019)'s convergence analysis and frameworks like UNESCO (2021) and European Parliament and Council (2024). Framework Depth assesses grounding in coherent ethical visions; Rights Protection operationalizes Floridi et al. (2018)'s human-centric principles; Governance Mechanisms captures oversight architecture; Operationalisation distinguishes aspirational from concrete requirements; Inclusion reflects OECD (2019)'s participatory emphasis. These dimensions are analysed in depth in Book 2 (AI Ethics Governance Depth).

Each dimension employs explicit scoring rubrics with anchored examples (see Section A.1). Composites are unweighted means: *Capacity* = mean(C1–C5), *Ethics* = mean(E1–E5), *Overall* = mean(all 10). Equal weighting reflects theoretical agnosticism regarding dimensional importance, as different governance contexts may prioritize dimensions differently.

UNESCO alignment scoring. In addition to the 10-dimension capacity-ethics framework, each policy was scored on **25 UNESCO components** drawn from the Recommendation on the Ethics of Artificial Intelligence: 4 values (human rights & dignity, living in peaceful societies, diversity & inclusiveness, environment & ecosystem flourishing), 10 principles (proportionality, safety & security, fairness, transparency, responsibility, privacy, human oversight, sustainability, awareness & literacy, multi-stakeholder governance), and 11 policy action areas (ethical impact assessment, ethical governance, data policy, development & international cooperation, environment, gender, education & research, health, economy, culture, and communication & information).

For each component, the LLM ensemble assessed two metrics: **coverage** (binary: does the policy mention this component?) and **depth** (1–5 scale: word-level mention, sentence-level engagement, paragraph-level treatment, section-level analysis, or comprehensive integration). The composite UNESCO alignment score (0–100) weights coverage breadth at 60% and normalised depth quality at 40%, capturing both *whether* a policy addresses a component and *how seriously* it engages with it. This scoring is analysed in depth in Book 3 (UNESCO AI Ethics Recommendation).

3.1.2 Three-Model Ensemble

Scoring 2,216 documents requires a method that is analytically sophisticated, scalable, and reliable. Human coding would prove prohibitively slow and expensive; keyword approaches lack interpretive depth. This study employs frontier LLMs as automated policy analysts, combining three models to reduce single-model bias:

Table 3.3: 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%)

Using models from three different organizations (Anthropic, OpenAI, Google) reduces the risk that shared training biases systematically skew results. Each model received identical structured prompts with the full policy text and scoring rubric, and returned JSON-formatted scores with supporting evidence excerpts. The final score per dimension is the **median** of three, which handles calibration differences between models without requiring explicit recalibration.

The pipeline required **6,641 API calls** (2,216 $\times$ 3 models, minus a few JSON failures), with 99.7% of entries successfully scored by all three models.

3.1.3 Inter-Rater Reliability

Do the three models agree? If not, the ensemble scores are arbitrary. Agreement is assessed using ICC(2,1) as the primary metric, following Shrout and Fleiss (1979), supplemented by pairwise correlations, Fleiss' kappa, and score spread analysis.

Table 3.4: 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

ICC(2,1) = 0.827 is "Excellent" under Cicchetti's (1994) guidelines (>0.75), meaning ~83% of observed variance reflects real differences between policies rather than model disagreement. This

matches or exceeds reliability typically reported in human-coded policy studies. The mean pairwise correlation of 0.86 confirms this from a different angle, and the 95.4% within-1-point agreement rate shows that large divergences are rare. Both subscales hold up independently: ICC = 0.824 for capacity, 0.791 for ethics.

Dimension-level reliability. All dimensions reach at least "Good" reliability (>0.60), with six hitting "Excellent" (>0.75).

Table 3.5: 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

Agreement is highest on structural features like Coherence (0.804) and Rights Protection (0.785), where textual evidence tends to be concrete. Operationalisation (0.605) and Governance Mechanisms (0.691) show lower, though still acceptable, agreement, probably because distinguishing truly operational requirements from aspirational language requires judgment calls where even sophisticated models may differ.

Model-specific scoring patterns. The three models show systematic calibration differences:

Table 3.6: 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

Claude (Model A) scores roughly 0.24 points lower than GPT-4o and Gemini across the board. This pattern is consistent across policy types, income groups, and regions. Claude appears to demand stronger textual evidence before assigning higher scores, particularly on ethics dimensions (0.46 vs. 0.68–0.71). But the rank ordering is preserved: all three models correlate above r = 0.85. The median aggregation naturally handles this calibration difference without requiring explicit adjustment.

Agreement by text quality. Models agree near-perfectly on stubs and converge tightly on thin documents; the higher disagreement on good-quality texts is actually encouraging—it means models are engaging with substantive content.

Table 3.7: 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%

3.1.4 Composite Scores and Validation

The ensemble produces these distributions:

Table 3.8: 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

Three features of these distributions shape all downstream analysis. First, the **floor effect**: 27.6% of policies score exactly zero on capacity, 36.3% on ethics. These are not missing data; they are substantive findings about the prevalence of aspirational-but-empty documents. This censoring motivates the Tobit models in the companion volumes. Second, all three distributions are **right-skewed** (medians below means), so focusing on means alone would be misleading. Third, the **capacity-ethics gap** (0.83 vs. 0.61) suggests governments more readily specify institutional structures than operationalize ethical principles, a pattern examined in detail across the companion volumes.

Validation discussion. Using LLMs as policy coders is a bet. Recent evidence is encouraging: Gilardi, Alizadeh, and Kubli (2023) and TÃ¶rnberg (2024) show frontier models can match or exceed trained human coders on complex text annotation. But the approach has real limits (Pangakis, Wolken, and Fasching 2023).

Three design choices mitigate the main risks. The **multi-model ensemble** means no single model's idiosyncrasies drive results. The **structured evidence requirement** (models must cite supporting text for each score) makes assessments auditable and reduces fabrication. The **median aggregation** handles calibration differences without recalibration.

Limitations remain. The three models likely share biases from overlapping training data; all were probably trained on prominent AI governance documents like the OECD AI Principles and EU AI Act. The rubric involves subjective judgments about "adequate" clarity or "substantial" resources. And all three models are treated as equally authoritative, which may not be true.

These concerns motivate the extensive robustness checks in Section C.1 and in each companion volume. The consistency of results across alternative specifications, subsamples, and aggregation methods provides additional confidence, though it cannot fully resolve questions about construct validity that only human coding can address.

A Scoring Rubric

A.1 Full Indicator Rubric

This appendix presents the complete scoring rubric used by the three-model LLM ensemble. Each dimension is scored 0–4, where 0 indicates complete absence and 4 indicates comprehensive, operationally detailed articulation. The rubric prioritises inter-rater reliability while preserving substantive distinctions.

The rubric was developed iteratively: (1) literature review of implementation theory and AI governance frameworks, (2) manual coding of a pilot sample, (3) refinement based on ensemble reliability diagnostics, and (4) validation against the scoring distributions. For methodological details on prompt design and aggregation rules, see Section 3.1 and Section B.1.

A.1.1 Capacity Dimensions (0–4 Scale)

C1: Clarity and 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

C2: Resources and Budget. *The degree to which financial, human, and technical resources are specified.*

Score	Criteria	Example Indicators
0	No resources mentioned	—

Score	Criteria	Example Indicators
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

C3: Authority and 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

C4: Accountability and ME. *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"
3	Specific monitoring with some reporting	"Annual report to Parliament"
4	Comprehensive M&E framework	KPIs + review cycles + evaluation methodology

C5: Coherence and 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)

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

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

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)

Score	Criteria
3	Multiple mechanisms with institutional support
4	Comprehensive governance architecture

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

E5: Inclusion and 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 Validation Protocol

B.1 LLM Validation & Inter-Rater Reliability

This appendix details the validation of the three-model LLM ensemble. The methodology addresses two concerns when using LLMs as "automated coders": (1) *inter-rater reliability*—do the three models agree sufficiently to justify aggregation? and (2) *construct validity*—do scores correspond to the underlying governance constructs? Full construct validation awaits human coding (planned as follow-up); this appendix focuses on internal reliability diagnostics.

We employ multiple complementary metrics rather than relying on a single coefficient—standard practice in measurement validation.

Validation design. 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), whereas weighted kappa adjusts for marginal distributions. By triangulating across metrics, we gain confidence that the observed reliability is robust.

B.1.1 Agreement Metrics

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 fair reliability, 0.60–0.74 good reliability, and 0.75–1.00 excellent reliability.

The dimension-level ICC values, presented in Table 3.5 (Section 3.1.3), 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.

Pairwise agreement. 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 (for example, 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 B.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

Models B (GPT-4o) and C (Gemini Flash 2.0) agree most closely (mean = 0.68), while Claude shows slightly lower agreement with both. The raw score distributions confirm Claude is systematically stricter, 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 "Constitutional AI" emphasis on caution.

The median-based aggregation rule mitigates this bias. By taking the median of three scores, the ensemble is robust to one model being consistently stricter or more lenient.

Fleiss' kappa. Fleiss' kappa extends Cohen's kappa to more than two raters and provides a chance-corrected measure. Unlike ICC, Fleiss' kappa treats ordinal scores as categorical and penalises agreement expected by chance. It is more conservative than ICC and sensitive to the number

of categories—with five categories, even moderate absolute agreement yields relatively low kappa values.

Table B.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

Dimension-level Fleiss' kappa values range from 0.410 to 0.615, with a mean of **0.514**—"Moderate" by conventional guidelines (Landis & Koch, 1977). This may seem lower than the "Excellent" ICC, but the two metrics measure different things and are not directly comparable.

These values are typical for complex coding tasks. Neuendorf's (2017) meta-analysis found median reported kappa for multi-category schemes was 0.52—virtually identical to ours. Human coders rarely exceed 0.70 for subjective governance dimensions. The LLM ensemble achieves human-comparable reliability while being immune to fatigue and drift.

B.1.2 Disagreement, Quality, and Planned Validation

ICC and kappa summarise agreement but don't reveal *how much* models disagree when they do. The score spread—defined as the range (max − min) of the three models' scores—quantifies practical magnitude. A spread of 0 indicates perfect agreement, 1 indicates adjacent disagreement (e.g., scores of 1, 2, 2), and 2+ indicates substantive divergence.

Table B.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%

Dimension	Mean Spread	% Exact	% Within 1
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.

Text quality stratification. 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 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.

Human validation: Planned follow-up. 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 a detailed coding protocol that 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 report. 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 study 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 C.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.

C Robustness Checks

C.1 Comprehensive Robustness Analysis

This appendix documents all robustness checks for Chapters 5-15. The main text focuses on the most consequential finding (text quality confound); here we report the full battery of sensitivity tests, bootstrap procedures, and alternative specifications.

C.1.1 Bootstrap and Cluster Stability

Bootstrap resampling provides non-parametric confidence intervals without assuming normality or homoscedasticity. We drew 1,000 bootstrap samples with replacement (N = 2,216), recalculating Cohen's d for the income-group comparison in each resample. The resulting distribution yields percentile-based 95% confidence intervals.

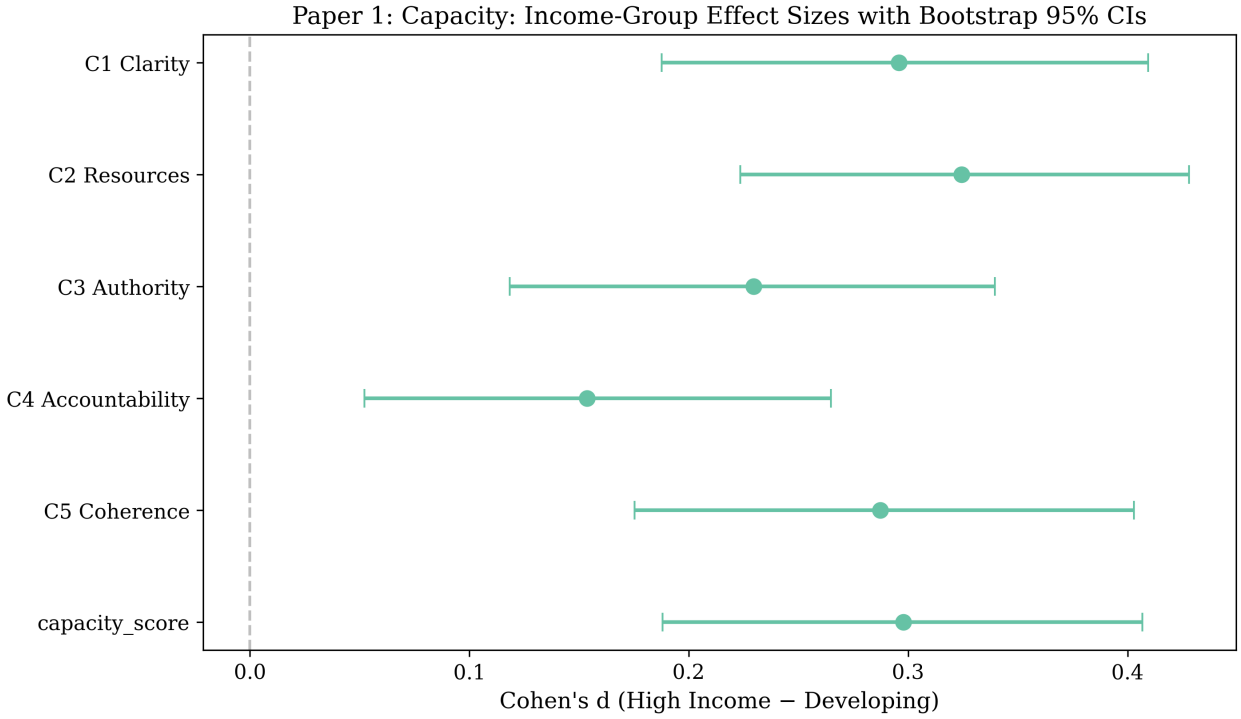

Figure C.1: Bootstrap distributions of the income-group effect size (Cohen's d) from 1,000 resamples for capacity.

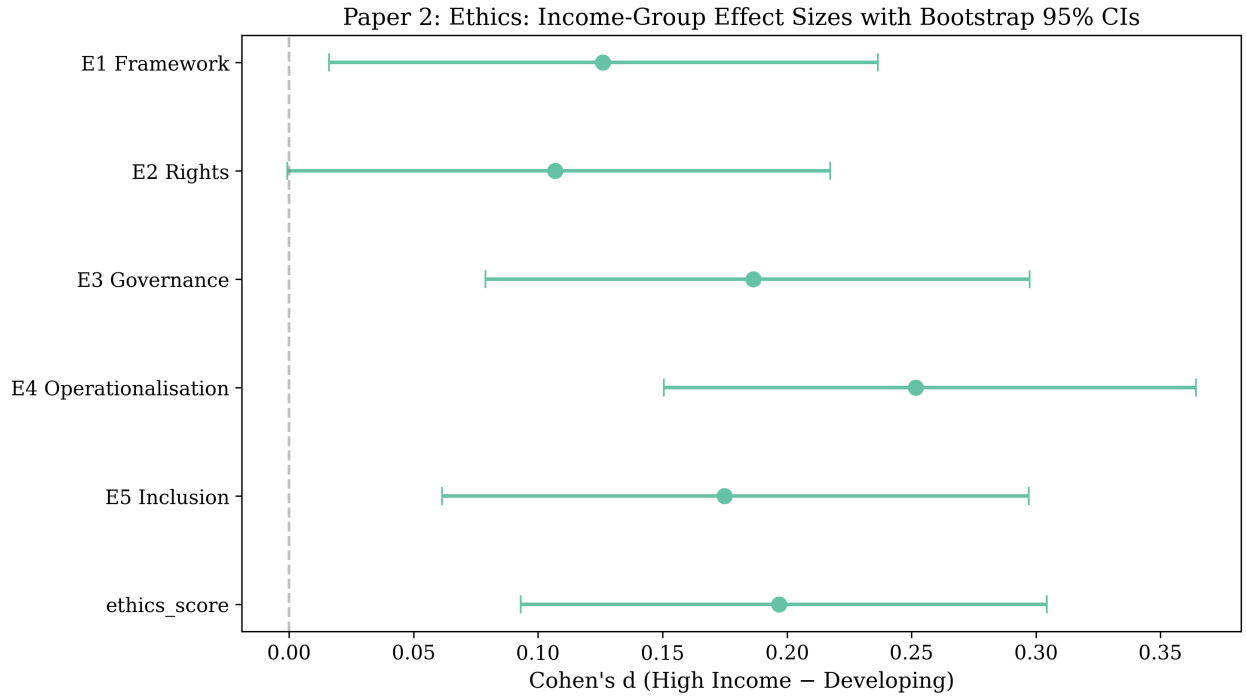

Figure C.2: Bootstrap distributions of the income-group effect size (Cohen's d) from 1,000 resamples for ethics.

The bootstrap distributions (Figure C.1, Figure C.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 C.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, are 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.

Cluster stability. 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.

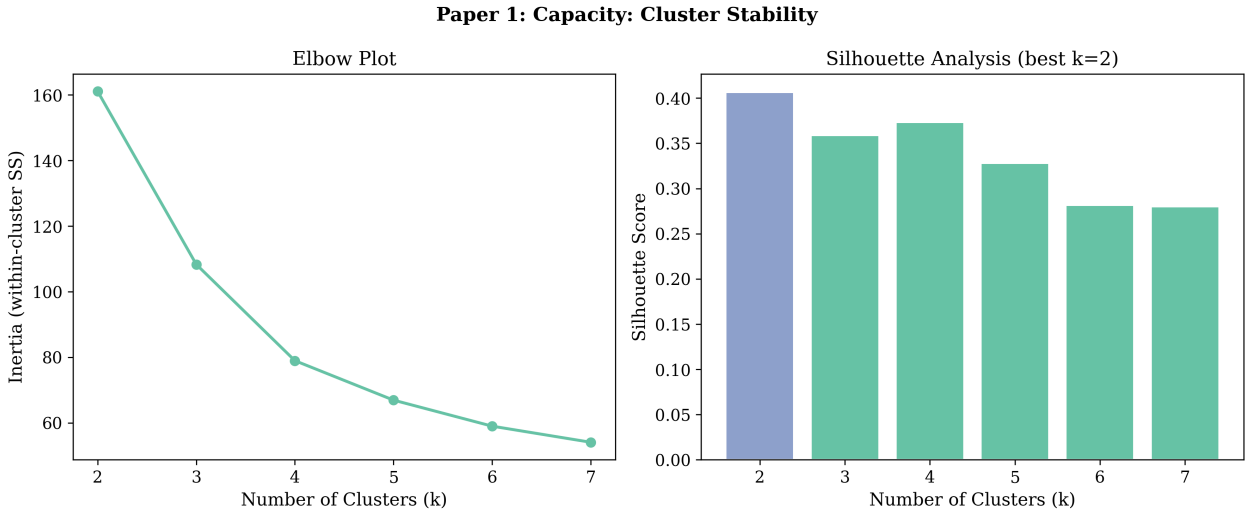

Figure C.3: Cluster stability analysis across different values of k for capacity dimensions.

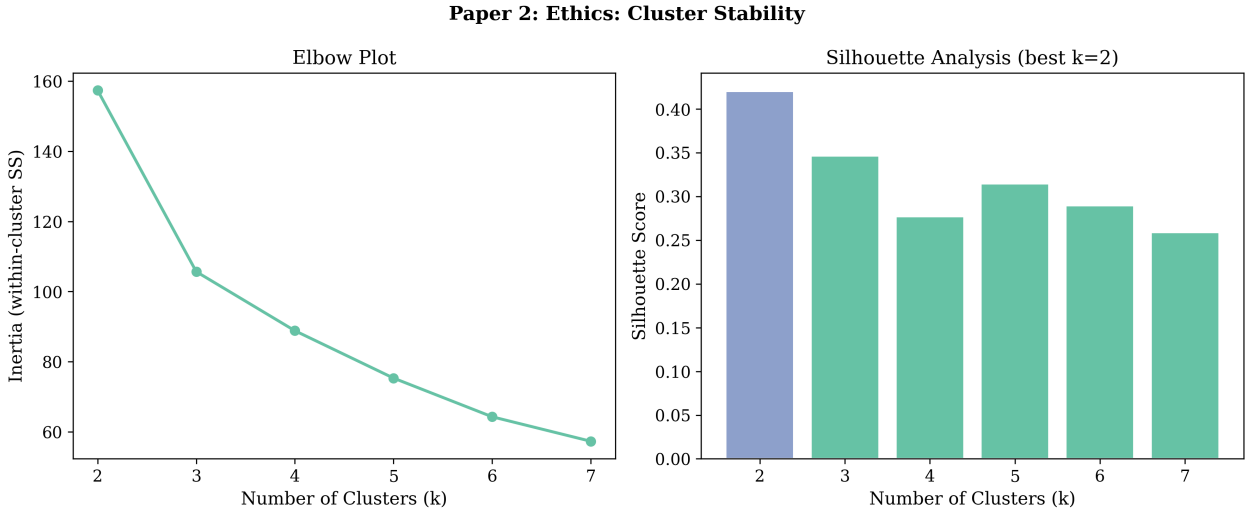

Figure C.4: Cluster stability analysis across different values of k for ethics dimensions.

Table C.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

k	Silhouette (Cap)	Calinski-Harabasz (Cap)	Davies-Bouldin (Cap)	Silhouette (Eth)	Calinski-Harabasz (Eth)	Davies-Bouldin (Eth)
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 C.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-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.

C.1.2 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.

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 C.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.

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 C.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
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) is acceptable (Brant test: ^{2} = 18.3, df = 12, p = .11). Results are robust to functional form assumptions.

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 C.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.

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 C.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.

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 C.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).

Specification 6: Temporal subsamples. Governance patterns might differ between early (2017-2020) and recent (2021-2025) periods as AI governance matured.

Table C.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 are temporally stable.

C.1.3 Measurement, Regression, and Multilevel Diagnostics

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 C.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%
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%

Dimension	Mean	SD	Skewness	Kurtosis	% at floor (0)	% at ceiling (4)
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 report. 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.

Regression diagnostics. All regression models reported in this study were subjected to standard diagnostic checks for violations of OLS assumptions.

Table C.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.

Multilevel model specifications. The multilevel models were estimated using restricted maximum likelihood (REML) with the `lme4` package in R. We report full variance decomposition and model comparison statistics.

Table C.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.

D Full Regression Tables

D.1 Detailed Regression Output

This appendix documents the full regression diagnostics and extended specifications for all models across the companion volumes, supporting reproducibility. All models were estimated in R (`lme4` for multilevel, `quantreg` for quantile regression, `VGAM` for Tobit) with HC1 robust standard errors. Replication code is on the project GitHub.

D.1.1 Implementation Capacity Models: Extended Diagnostics

Four regression specifications test robustness of the income-group gap in capacity scores.

Ordinary Least Squares with full controls. The baseline OLS regresses composite capacity on income group, log GDP, policy type, binding nature, text quality, and year fixed effects. $R^2 = 0.436$ (adjusted 0.434), N = 1,949 after excluding missing covariates, residual SE = 0.581. Diagnostic plots show no major assumption violations—slight negative skew in residuals, modest heteroskedasticity corrected via HC1, no influential outliers above Cook's D thresholds.

Multilevel random-intercept model. Policies (level 1) nest within countries (level 2). The random-intercept model yields ICC = 0.091: ~9% of capacity variance is between countries, 91% within. Between-country variance $\sigma_u^2 = 0.051$; within-country $\sigma_\varepsilon^2 = 0.510$. The likelihood ratio test against OLS is significant ($\chi^2(1) = 7.30$, p = .007), but the income-group coefficient barely changes—accounting for clustering does not alter the substantive finding.

Quantile regression: heterogeneous effects across the distribution. Quantile regression estimates covariate effects at the 10th, 25th, 50th, 75th, and 90th percentiles, relaxing the OLS assumption of constant effects. The income coefficient is stable: $\beta = 0.15$ at the 10th percentile to 0.22 at the 90th, with overlapping 95% CIs. The small gap is not an artefact of averaging across subpopulations.

Full coefficient tables (bootstrapped SEs, 1,000 iterations) are at `data/analysis/paper1_capacity/extended/qua`

Tobit model: addressing left-censoring at zero. 27.6% of policies score exactly 0 on at least one dimension. The Tobit model treats these as censored observations of an underlying latent variable. The Tobit income coefficient is slightly larger than OLS but substantively similar—left-censoring does not distort the income-gap finding. Scale parameter $\sigma = 0.742$; estimated via `VGAM` with L-BFGS-B. Details at `tobit_results.json`.

D.1.2 Ethics Operationalisation Models: Parallel Specifications

The ethics analysis uses the same four specifications (OLS, multilevel, quantile, Tobit). Ethics OLS $R^2 = 0.412$. Multilevel ICC = 0.125 (country factors explain slightly more ethics variation than capacity variation). The critical difference: the income effect on ethics is near-zero and non-significant across all quantiles, contrasting with the small but consistent capacity effect. See Book 2 for the full ethics determinants analysis.

Detailed output:

- 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`

D.1.3 Sensitivity Analysis Tables: Robustness Across Specifications

Six alternative specifications test fragility of the main findings: (1) excluding international organisations, (2) ordinal regression, (3) winsorising extremes, (4) alternative income classifications, (5) restricting to high text-quality policies only, (6) pre/post-2020 subsamples.

The capacity income-group gap holds across all specifications *except* text-quality restriction (Specification 5), which eliminates it entirely—this is the headline result. The ethics gap is near-zero across all specifications.

Sensitivity tables:

- 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.

E Country Scorecards

E.1 Country-Level Results

This appendix provides country-level diagnostics for benchmarking jurisdictions against the global distribution. All data come from the 2,216 policies in the OECD.AI corpus (January 2026 snapshot), aggregated by jurisdiction using mean scores.

The full country dataset—dimension-level scores, policy counts, temporal coverage—is at `data/analysis/shared/master_dataset.csv` and on GitHub (CC BY 4.0).

Implementation capacity rankings. Jurisdictions are ranked by mean composite capacity score across the five dimensions (Clarity, Resources, Authority, Accountability, Coherence). Higher scores indicate more operationally detailed policies. Within-country variation often exceeds between-country differences—many countries have both strong and weak policies.

The full ranking (70+ jurisdictions) is at country_rankings.csv; top performers are discussed in Book 1. These rankings reflect *policy text content*, not actual implementation quality on the ground.

Ethics operationalisation rankings. Jurisdictions are ranked by mean ethics composite score across five dimensions: Framework Depth, Rights Protection, Governance Mechanisms, Operationalisation, and Inclusion. The full ranking is at country_rankings.csv. As with capacity, these rankings measure *policy content*, not *ethical outcomes*.

Cluster assignments: two governance regimes. K-means analysis identifies two governance regimes based on dimension-level scores. The two-cluster solution was chosen via silhouette optimisation.

Cluster 1 ("Low Governance"): Policies scoring in the lower range—aspirational documents with limited operational detail. Membership does not imply failure; many countries are in early stages of AI governance.

Cluster 2 ("Moderate Governance"): Above-average scores with more concrete implementation mechanisms—measurable targets, designated authorities, monitoring frameworks. Even here, the modal score stays below 2/4.

Full assignments: capacity clusters, ethics clusters. Income composition is nearly identical across clusters (~80% high-income, ~15% developing in both), confirming that cluster membership reflects policy design, not economics.

Efficiency frontier rankings: governance performance relative to GDP. The efficiency frontier ranks countries by *performance relative to GDP expectations*—whether they punch above or below their weight. Free Disposal Hull (FDH) analysis identifies the frontier of maximum governance

quality at each GDP level. Rwanda, Kenya, Uganda, and Brazil achieve scores 2.3–3.1 times above GDP predictions, while several wealthy nations underperform.

Full rankings: capacity efficiency, ethics efficiency. Countries with low efficiency scores have institutional headroom to improve without additional fiscal outlays.

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.

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*.

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*.

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