



**Flinders
University**



Jeff Bleich Centre
for Democracy and
Disruptive Technologies

Sovereign AI Through Governance Harnesses: Evidence from a Multi- Jurisdiction Evaluation

Ben Kereopa-Yorke, Rodrigo Praino

AUGUST 2026



© Flinders University, 2026

Except where noted, this work is licensed under CC BY 4.0

Images sourced under licence from Canva are excluded from the CC BY licence.

Suggested Citation: Kereopa-Yorke, B., Praino, R. (2026). "Sovereign AI Through Governance Harnesses: Evidence from a Multi-Jurisdiction Evaluation", Jeff Bleich Centre for Democracy and Disruptive Technologies, Flinders University, South Australia
DOI: <https://doi.org/10.25957/79y6-wb39>

jbc@flinders.edu.au

Sovereign AI Through Governance Harnesses: Evidence from a Multi-Jurisdiction Evaluation

Ben Kereopa-Yorke¹

Rodrigo Praino²

Abstract

Sub-national governments are increasingly told that genuine AI sovereignty requires procuring or building expensive frontier models under domestic control. This framing conflates two distinct properties: the capacity to access AI capability, and the autonomy to govern its use. We argue, and demonstrate empirically, that for the public-sector advisory tasks that dominate everyday demand (housing, health navigation, legal information, social services), the autonomy property is best supplied not by a sovereign model but by a separable governance layer, which we call a governance harness. The harness wraps any frontier model in five auditable components: a policy router, a source-authority gate, a citation-enforcement contract, an escalation/refusal layer, and a jurisdiction-specific governance prompt. We evaluated this layer across four frontier models (Claude Opus 4.7, Claude Sonnet 4.6, GPT-5.5, GPT-5.4-mini), nine experimental conditions, two Australian jurisdictions (South Australia, $n=200$; Victoria, $n=50$), and five component ablations, producing approximately 12,600 scored responses at a total compute cost of approximately USD 181. The harness produced statistically significant improvements for every model (Opus +22.2pp; Sonnet +15.6pp; GPT-5.5 +9.4pp; GPT-5.4-mini +5.5pp), with the largest deltas concentrated on the models that started from the lowest baseline scores on this benchmark ($r=-0.988$ between baseline and delta). We caveat this pattern with a ceiling-effect interpretation: stronger baselines have less remaining headroom to gain, and our highest scorer reached only ~72%, well short of the 100% ceiling. Citation enforcement and source-authority verification together explain a range of the harness effect depending on model, ranging from 39% to 97% of the total harness effect, with one model (GPT-5.4-mini) showing super-additive component interactions, identifying these two components as the highest-value governance investment, with the important caveat that the leave-one-out design assumes roughly additive component effects and at least one model exhibits clear interaction. The cheapest model tested (GPT-5.4-mini), which also had the highest baseline score on this benchmark, achieved the best governed outcome at a fraction of the cost of larger models (roughly 1/22 the per-answer cost of the most expensive model evaluated). We discuss the durability of this pattern and its sensitivity to provider list prices in Section 5.2. Effect sizes transferred to Victoria with zero code changes ($n=50$, two models). We argue, with explicit limitations, that sovereign AI value for sub-national governments is principally a governance-engineering investment, not a model-procurement decision.

¹ Ben Kereopa-Yorke (benke@microsoft.com) is Research Fellow at the Jeff Bleich Centre for Democracy and Disruptive Technologies at Flinders University and Senior Offensive Security Researcher at Microsoft.

² Rodrigo Praino (rodrigo.praino@flinders.edu.au) is Professor of Politics and Public Policy and Director of the Jeff Bleich Centre for Democracy and Disruptive Technologies at Flinders University.

1. Introduction

Sub-national governments (Australian states, Canadian provinces, US states, European regions) confront a sovereignty dilemma in artificial intelligence. They depend on a small number of externally controlled frontier model providers to deliver capability that citizens and public servants increasingly expect, yet the use cases that matter most for them (housing assistance, health navigation, legal information, social services) demand outputs grounded in locally legitimate sources, appropriately escalated when risk is high, and structured for institutional accountability. A common policy response to this tension has been to commission or procure a 'sovereign model'. This framing is intuitive but, we will argue, mis-specified for the advisory use cases that dominate sub-national workloads. It conflates the capacity to access AI capability with the autonomy to govern its use, and consequently directs investment toward the property that sub-national governments are least able to supply (model construction) rather than the one they are most able to supply (governance engineering).

The evaluation literature compounds this confusion. Public LLM benchmarks measure capability (accuracy on examinations, reasoning on puzzles, fluency on open-ended tasks) but are silent on the properties that actually determine whether a model is fit for public-sector advisory use: does the output cite a legitimate local authority; is that source the current version; is a risky query routed to a human; can a downstream auditor trace the chain of evidence? To our knowledge, no established methodology quantifies what a governance layer adds, or which of its components carry the value. As a result, jurisdictions weighing sovereign-AI investments have no empirical basis on which to choose between building a model and building a harness.

This concept note presents a methodology, a prototype system, and a multi-jurisdiction evaluation that together aim to close that gap. The prototype, which we call a governance harness, is a thin, auditable software layer that sits between users and one or more frontier models and enforces five named functions: policy routing, source-authority gating, citation enforcement, escalation/refusal, and a jurisdiction-specific governance prompt. The harness is neither a model nor merely retrieval-augmented generation (RAG). It converts a capable model into an accountable one. We evaluated this layer across four frontier models, nine experimental conditions, two Australian jurisdictions, and five component ablations, producing approximately 12,600 scored responses.

Six findings, developed in sequence in Section 4, organise the argument:

1. The harness effect is universal: all four models improved with statistically significant deltas between model-only and harness conditions (Section 4.1; Table 1; Figure 3).
2. The effect is largest for the lowest-scoring baselines: harness delta is strongly negatively correlated with baseline score on this benchmark ($r=-0.988$; Section 4.2; Figure 4). We discuss the role of ceiling effects in this pattern in Section 4.2.
3. Two components, citation enforcement and source authority, supply most of the gain, explaining 39% to 97% of the harness effect depending on model (Section 4.3; Table 2; Figure 5). This is the mechanism: governance value comes from binding outputs to legitimate, audited evidence.
4. Effect sizes transferred to Victoria with config-only changes, indicating the methodology is portable rather than SA-specific (Section 4.4; Table 3; Figure 6).

5. Cost and capability rankings dissociated under the harness in this test set: the cheapest model tested produced the highest harness score, while the most expensive produced the lowest (Section 4.5; Table 4; Figure 7).
6. The leakage gate is clear: the harness effect is not an artefact of metadata cues (Section 4.6; Table 5; Figure 8).

We make a bounded sovereignty claim. We do not show that sub-national governments can replace frontier providers. We show that, with a small auditable layer, they can convert externally controlled capability into locally accountable use. The harness does not deliver independence from providers. It delivers governance over them. This is a narrower form of sovereignty than the term sometimes implies, but it is one that sub-national governments can actually build, audit, and operate, and on the evidence reported here it is also the form that yields most of the practical value at a small fraction of the cost.

2. Theoretical Framework

2.1 A multidimensional framework for AI sovereignty

With the rise and rapid development of artificial intelligence (AI) seen in recent years, governments around the world have begun to embrace the idea of “sovereign AI”, or the ability to design, deploy and govern AI and all of its components (Aaronson, 2025). The term has migrated, with remarkable speed, from industry into national strategies, defence reviews and sub-national procurement plans. Yet an interpretive framework for what AI sovereignty actually means, and how it is achieved, is still largely missing. In fact, AI sovereignty remains an ill-defined concept, which is invoked to mean conceptually different things such as a state’s regulatory power or its strategic autonomy (see Hoeffler & Mérand, 2024; Krasner, 1999; Krishnamurthy, 2025). In the pages that follow we sketch a comprehensive AI sovereignty framework by adapting a heuristic tool originally developed by Cappelli (2010) and subsequently operationalised and applied by Aliberti, Cappelli and Praino (2023).

2.1.1 The unidimensional trap: AI sovereignty as capacity

Up to this moment, the conversation about AI sovereignty has developed essentially unidimensionally, centred on the idea of AI capacity. Most of this debate has been centred on the ability to build AI infrastructure, stand up data centres, secure enough computing power, lock in a supply chain of the products and inputs necessary for AI (e.g., advanced semiconductors, high-bandwidth memory, networking, energy, data), train domestic talent, and, ultimately, own a frontier model. “Sovereign AI,” according to this framing, is something a jurisdiction has once it possesses enough of this material stock. As Srivastava and Bullock (2024) point out, this view is reflected in industry usage, with Nvidia, for instance, defining sovereign AI as “a nation’s capabilities to produce artificial intelligence using its own infrastructure, data, workforce and business networks”. Clearly, this definition equates sovereignty with productive capacity.

This is the same trap that Aliberti, Cappelli and Praino (2023), analysing a different sector, call the “tyranny of capacity.” Capacity is the dimension that is most readily seen, counted and communicated. Because it can be proxied by objective indicators (e.g., the number of accelerators, megawatts of data-centre power, parameter counts, or dollars of investment) it tends to swell until it stands in for the entire

concept, an instance of the conceptual overstretch against which Sartori (1970) warned. In the process it eclipses, and ultimately hides, a second dimension that is no less constitutive of sovereignty: autonomy.

The cost of this capacity-only approach to AI sovereignty is not merely academic. In fact, a capacity-only account of AI sovereignty cannot distinguish between an actor that owns its AI stack and governs it on its own terms, and an actor that owns an impressive stack whose rules of use, model behaviour, content policies and update cycles are in fact determined elsewhere by a foreign provider, a foreign regulator, or a domestic corporate interest powerful enough to set the terms (see Srivastava, 2022, 2023). The first is sovereign; the second is capable but not sovereign. A framework that measures only capacity is structurally blind to this difference. Without this distinction being explicitly made, even ostensibly sovereign national AI stacks remain dependent on transnational cloud infrastructure, cross-border data flows, and multinational hardware supply chains (Ishkhanyan, 2025). Interestingly, the narrowness of this capacity-only framing is compounded by the ultimate concentration of the underlying capacity. In fact, only a small handful of firms supply some of the critical technology needed to boost capacity, with Nvidia alone accounting for roughly 92% of data-centre GPUs and TSMC for around 65% of global foundry output (Weymouth, 2025). As a consequence, the pursuit of capacity, even when waged as “AI nationalism” (Aaronson, 2025), tends to leave significant external dependencies in place, given the reliance on a small number of suppliers that a capacity-only strategy entails. This shows how even when capacity grows systematically, autonomy can still remain out of reach.

2.1.2 From capacity to sovereignty-as-status: the two-dimensional concept

The corrective is to carry into the AI domain a distinction that Aliberti, Cappelli and Praino (2023) developed in another setting: that between sovereignty as a set of means and resources and sovereignty as a status, that is, the difference between being merely capable and being genuinely sovereign. Building on their account, we treat AI sovereignty not as a single property a jurisdiction simply has or lacks, but as a composite one, generated by the interplay of two distinct dimensions, each necessary and neither sufficient on its own:

- Capacity: the material ability to access, build, deploy and operate AI, and to use it to achieve goals across the economic, administrative, social, scientific and security realms.
- Autonomy: the decision-making ability to determine, independently, what an actor wants AI to do and how it should behave, and to govern its use under domestic norms and accountability mechanisms, free from coercion or imposition by divergent interests, whether foreign or domestic.

Capacity is achieved through building infrastructure, securing supply chains, obtaining computing power, and obtaining the most powerful models. Autonomy is achieved by regulating, implementing and controlling model outputs and, as we argue below, can be exercised through a separable governance layer even over a model one does not own. Both dimensions are required for the status of AI sovereignty. An actor with high capacity but low autonomy, or high autonomy but low capacity, occupies a different and lesser status, being either capable or self-reliant, but not sovereign in the full sense. The same fault line runs through the wider literature, which separates sovereignty exercised *through* AI, that is, competence built on (often privately owned) infrastructure, from sovereignty exercised *over* AI, that is, institutional control of how it is used (Srivastava & Bullock, 2024). The former is capacity, the latter autonomy.

Crucially, autonomy is the *sine qua non* of sovereignty-as-status: no actor can count as fully sovereign unless it can set its own objectives and decide for itself how to pursue them. In the AI context, this should not be read as Westphalian sovereignty, but rather as the ability to make decisions over the contested choices that are properly the jurisdiction's own. In other words, choices that sit above a thin safety baseline that no responsible actor would itself wish to remove. That baseline is akin to *jus cogens*, or a floor of constraints that bound every sovereign, regardless of whether or not they authored these rules.

2.1.3 The capacity dimension

Capacity is the material ability to access, build, deploy and operate AI. It runs from the tangible assets that the current debate already captures well (e.g., compute, including accelerators and the data centres, power and networking that house them; access to and control over data; the supply chain for critical inputs; the energy base; the pool of technical talent; and the models themselves) through to applied AI, or the ability to effectively utilise and integrate that capability within national (or sub-national) policy, infrastructure, public administration, the economy and society. Capacity is therefore not only what an actor owns but how well it uses it, including absorptive capacity, institutional uptake, diffusion across sectors, the quality of public-service delivery built on AI, and the embedding of AI into welfare, health, education, the environment and economic policy rather than only defence and security. An actor can be rich in assets yet poor in application, owning the kit but failing to convert it into effective policy and service outcomes. The most visible and fundable part of capacity is the material stock, because it lends itself to objective indicators, but capacity in full is measured just as much by how deeply AI is woven into the fabric of public and economic life. In addition, for most jurisdictions, capacity is always going to be only partially attainable. In fact, even powerful states will have to rely on partnerships to assemble the necessary compute, talent, energy and capital that AI requires to continue its journey (Aaronson, 2025; Weymouth, 2025).

2.1.4 The autonomy dimension

Autonomy is the decision-making ability to determine, independently, what an actor wants AI to do and how it should behave, and to govern its use under domestic norms free from coercion or imposition by divergent interests, whether foreign or domestic, above and beyond the shared values of a common safety baseline.

One part of autonomy is non-dependence, or the ability to access and operate critical AI without relying on external sources of supply, and so to ensure mission independence in case of necessity. The relevant test is not everyday self-reliance but resilience under disruption (e.g., export-control regimes, sanctions, price shocks, a provider's withdrawal or unilateral change of terms, geopolitical ruptures). An actor with high autonomy can keep its essential AI-enabled functions running when the external stack is withdrawn or otherwise disrupted. This does not imply autarky: a fully autonomous actor is not an isolationist one that builds everything itself, but one that can rely on diversified, substitutable sources of supply and fall back on domestic solutions when it must. This concept reframes part of the "sovereign stack" debate, in the sense that the goal is not to reproduce the entire stack domestically, but to eliminate single points of external control over functions that cannot be allowed to fail (see Calzada, 2025).

The other part of autonomy is the ability of decision-makers to formulate AI strategy, set the rules that govern model behaviour and outputs, and regulate AI in the public interest, however they define it. In the

public-service setting at issue here, this rule-setting is exercised by adding governance constraints on top of a provider's outputs (e.g., restricting eligible sources, requiring citations, and escalating or refusing high-risk queries) rather than by relaxing, overriding, or circumventing the safety policies the provider itself enforces; those policies remain in force regardless of how a jurisdiction configures the harness, which exposes no setting capable of disabling them. While regulation and governance can be set and implemented by decision-makers, empirically, there needs to be an instrument that allows the implementation of governance and regulatory decisions. This is where the harness discussed below comes into play.

This restriction is consistent with, rather than a departure from, autonomy as defined above. In fact, jurisdictions govern above a floor of constraints that represent a near-universal safety baseline (e.g., declining to provide instructions on how to build biological weapons, or to generate child sexual abuse material) that almost any responsible actor would adopt of its own accord. Such a floor is not a “divergent interest” imposed by a rival authority, but a shared background within which self-government is exercised; the value-laden, genuinely contestable choices (e.g., which sources count as authoritative, what is refused or escalated, or the epistemic posture the system adopts) all lie in the envelope above that floor, and over that envelope the harness gives the jurisdiction full control. This is also why the conservative design of the harness reinforces autonomy rather than limiting it. The harness can only ever add accountability and never subtract it, no actor, foreign or domestic, can use it to push behaviour below the floor.

Together, these two parts track a distinction long present in the digital-sovereignty literature, which treats the concept as comprising both regulatory power and strategic autonomy (Krishnamurthy, 2025). The non-dependence component is increasingly explicit in state doctrine. As Srivastava and Bullock (2024) point out, France's AI Commission warns that falling behind undermines their sovereignty, because it implies “a one-way dependence on other countries” that erodes the collective ability to make choices aligned with national values and interests. In the meantime, the European Union has reframed the same concern as “open strategic autonomy” (Aaronson, 2025; Sheng, 2022). Autonomy in this sense can also be pursued below the national level: the German state of Schleswig-Holstein decided to move away from proprietary Microsoft software to open-source alternatives explicitly as a way to enhance its digital sovereignty by reducing reliance on a foreign vendor (see Krishnamurthy, 2025).

2.1.5 The AI sovereignty matrix

Plotting the two dimensions against each other produces a simple matrix, with capacity on one axis and autonomy on the other. Its point is that AI sovereignty is not a single property an actor either has or lacks, but the product of two dimensions that can be present in different combinations (see Srivastava, 2017):

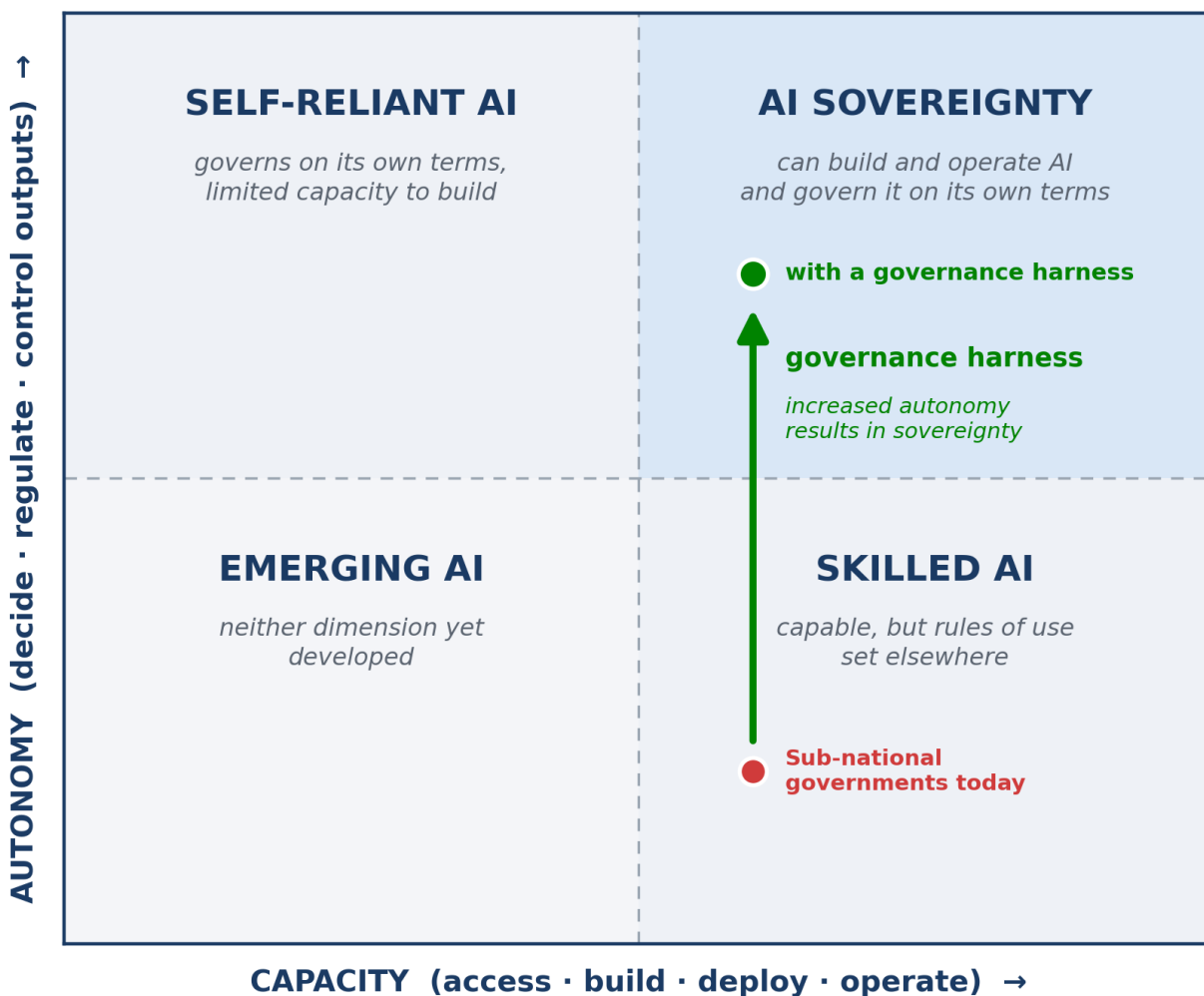
- **AI sovereignty:** high capacity and high autonomy. Sovereign AI takes place when an actor can build, deploy and operate AI capacity and govern its use on its own terms. Sovereignty is what the two dimensions produce when both are high, and is never delivered by capacity alone.
- **Skilled AI:** high capacity, low autonomy. Commanding or readily accessing substantial AI capability, while also being subject to rules governing how that capability is used that are effectively set elsewhere, by a foreign stack on which it depends, or by domestic interests that

Sovereign AI Through Governance Harnesses

dominate the process, results in Skilled AI that is capable and probably even useful, but not sovereign.

- **Self-reliant AI:** low capacity, high autonomy. Choosing to govern AI on one's own terms and secure non-dependence in critical functions will almost certainly result in low levels of capacity, barring ability to build your own frontier capability. This results in AI that is highly and tightly controlled, but that is likely to ultimately have limited means.
- **Emerging AI:** low capacity and low autonomy. This takes place when actors have yet to develop either dimension to a meaningful degree.

The AI Sovereignty Matrix



The matrix lets us position actors and follow how they shift over time. It makes clear that there is no single route to AI sovereignty: an actor can progress horizontally, by accumulating capacity, or vertically, by developing autonomy. The first can largely be bought; the second has to be built, through deliberate strategic investment in the conditions for AI autonomy.

This is exactly the situation of the sub-national governments at the centre of this note. Australian states, Canadian provinces, US states and European regions sit, characteristically, in the Skilled AI quadrant: they enjoy ready access to frontier capability through commercial APIs, and can invest in additional infrastructure to increase capacity and move right along the matrix, though for the advisory use cases at the centre of this note such capacity investment, including the construction of a domestic frontier model, is neither sufficient for sovereignty nor a precondition for it. But in order to actually achieve AI sovereignty *in lieu* of simply acquiring AI capabilities, they need to be able to actually exercise autonomy over how those models behave in public-service delivery. Sovereignty layered in this way can also assist in achieving “digital federalism” that distributes AI authority across local, national and global tiers rather than concentrating it at a single level (see Ishkhanyan, 2025).

2.1.6 From framework to instrument: the governance harness

This is where the framework meets the empirical contribution of this note. The governance harness described in the sections that follow is best understood as an instrument of autonomy, or the concrete mechanism by which an actor can climb the autonomy dimension that matters most for everyday public-sector use, in addition to spending resources to build capacity.

To a certain degree, the harness operationalises autonomy. It sets the rules by encoding jurisdictional norms at five auditable control points — a policy router, a source-authority gate, a citation-enforcement contract, an escalation/refusal layer and a jurisdiction-specific governance prompt — each a place where the rules governing model outputs are set, audited and contested locally rather than inherited from the provider. This is precisely “controlling model outputs”: regulating, implementing and enforcing how a model may speak in a housing, health, legal or welfare context, on the using jurisdiction’s terms. Because each of these controls can only narrow, qualify, or decline what the model would otherwise produce, the autonomy the harness supplies is designed to be conservative: it exposes no mechanism by which a jurisdiction could weaken, disable, bypass, or override the provider’s enforced safety policies, which remain fully in force at all times; the harness can only add accountability on top of them, never subtract it. The harness also reduces dependence: because it is model-agnostic and sits above any provider, it lessens single-provider reliance and curates a domestically controlled corpus of legitimate sources, so that the governance layer and the institutional knowledge embedded in it survive a change of model or provider. In doing so the harness instantiates, at the scale of a single jurisdiction, the broader insight that sovereignty over AI is exercised above all through rule-making over use rather than through ownership of the model (Srivastava & Bullock, 2024; Veale et al., 2023).

In the language of the matrix, a well-engineered harness moves a sub-national government vertically, that is, out of the Skilled AI quadrant and toward AI sovereignty, at a tiny fraction of the cost of keeping moving horizontally and ultimately failing to be sovereign anyway.

2.2 Harnesses as instruments of autonomy

We define a governance harness as a deterministic software layer, sitting between the user and one or more frontier models, that enforces five named functions. The harness is the operationalisation of the autonomy property: each function is a place where jurisdictional norms are encoded, audited, and contested. The five components are:

- Policy router (1) - classifies the incoming query by domain (housing, health, legal, civic, education and others) and risk tier, then selects the relevant source pool and prompt template.
- Source-authority gate (2) - restricts retrieval to a curated set of legitimate local authorities (state government, federal government, vetted NGO and health-information sources). Sources outside the gate are excluded from the model's context window before generation.
- Citation enforcement (3) - a structured output contract that requires each substantive claim to be tied to a retrieved source with sufficient precision for downstream audit.
- Escalation and refusal (4) - domain-and-risk rules that route specific query categories (acute health, legal advice on individual matters, crisis welfare) to a refusal or human-handover template rather than a generated answer.
- Governance system prompt (5) - a jurisdiction-specific system instruction that establishes epistemic posture: prefer the curated corpus, decline beyond it, express uncertainty explicitly.

2.3 How a harness differs from generic RAG

Retrieval-augmented generation (RAG) is now a mature pattern: a user query is embedded, similar passages are retrieved from a vector index, and the passages are concatenated into the model's prompt. RAG improves factual grounding but does not, in itself, perform governance work. It does not police which sources are eligible to enter the context; it does not require the model to bind its output to retrieved citations; it does not route risky queries away from generation; and it does not impose a jurisdiction-specific epistemic posture. Figure 1 contrasts the two architectures.

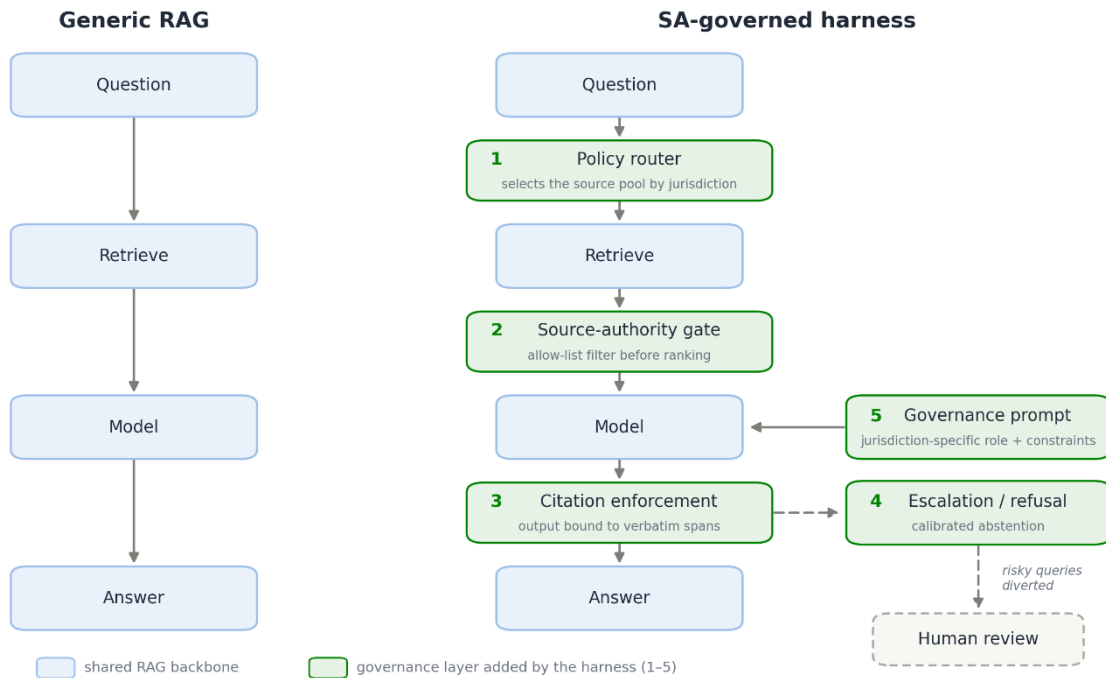


Figure 1. Generic RAG (left) versus governance harness (right). Both share a retrieval step against a corpus. The harness adds a policy router that selects the source pool, a source-authority gate that filters retrieval candidates by allow-list before ranking, a citation-enforcement contract on the output, an escalation layer that can divert risky queries to a human, and a jurisdiction-specific governance prompt.

The remainder of this note quantifies what each of these governance functions is worth, in isolation and in combination, and asks whether the resulting effects justify treating the harness rather than the model as the primary unit of sovereign-AI investment.

3. Methods

Section 3.1 describes the source corpora that the source-authority gate and retrieval layer draw from. Section 3.2 describes the benchmark items used to probe the system. Section 3.3 enumerates the nine experimental conditions. Sections 3.4-3.6 describe the ablation design, the cross-jurisdiction protocol, and the scoring methodology.

3.1 Data curation

Two jurisdictional corpora were constructed using an identical pipeline. The pipeline starts from a seed list of named state-government, federal-government, and vetted non-government authorities. Discovery extends the seed via site-internal link traversal, filtered by domain allow-lists. Ingestion captures full text, metadata, and last-modified timestamps. Curation drops superseded, duplicated and non-textual content, then applies a per-domain authority filter. Surviving documents are chunked and embedded into a vector index used by both the retrieval layer and the source-authority gate. The South Australian corpus reached 951 curated sources; the Victorian corpus 913. The Victorian corpus was built later, using the same pipeline driven by a different JurisdictionConfig (see Section 3.5). Figure 2 summarises the pipeline.

One pipeline builds each jurisdiction's corpus

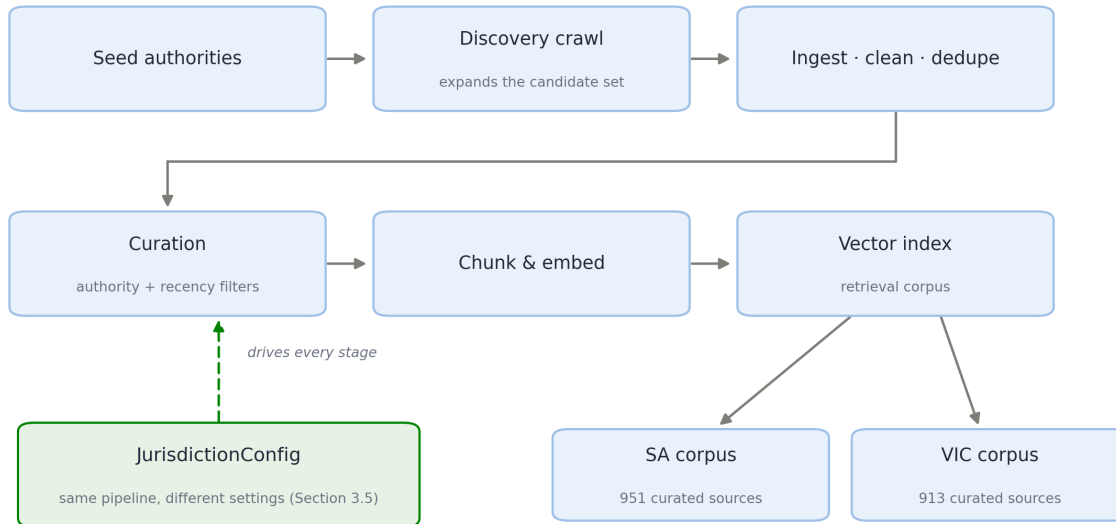


Figure 2. Source-curation pipeline. Seed authorities are expanded by discovery crawling; candidate sources are ingested, cleaned and deduplicated; a curation step applies authority and recency filters; the resulting corpus is chunked and embedded into a vector index. The South Australian corpus contains 951 curated sources; the Victorian corpus contains 913.

3.2 Benchmark design

The South Australian benchmark contains 200 items stratified by domain, risk level, and difficulty. Items were authored by the research team to reflect plausible public-sector advisory queries, with each item carrying a known-correct authoritative source, an expected risk tier (low/medium/high), and a difficulty rating. Table M1 summarises the domain distribution.

Domain	Items	Risk-tier mix
Housing & tenancy	30	low-high
Health navigation	25	low-high
Legal information	25	medium-high
Civic / government services	20	low-medium
Education & training	15	low-medium
Welfare & social services	25	medium-high
Employment & workplace	20	low-medium
Cross-domain / other	40	low-high
Total	200	-

Table M1. Domain composition of the South Australian benchmark (n=200).

The Victorian benchmark contains 50 items, constructed with the same authoring protocol against Victorian sources. Its smaller size reflects its role as a transfer pilot rather than a full evaluation.

3.3 Experimental conditions

Each benchmark item was run under nine experimental conditions, designed to isolate the contributions of retrieval, governance, and metadata. Table M2 enumerates the conditions and what each isolates.

#	Condition	What it tests
1	Model-only	Baseline capability without retrieval or governance.
2	Generic RAG	Retrieval contribution without governance.
3	SA harness	Full harness: policy router + authority gate + citations + escalation + prompt.
4	Metadata-minimal	Provides only document titles and URLs; tests metadata leakage.
5	Compressed-retrieval fusion	Harness combined with summarised retrieval context.
6	Ablation: no citation enforcement	Removes component 3.
7	Ablation: no source-authority gate	Removes component 2.
8	Ablation: no governance prompt	Removes component 5.
9	Ablation: no escalation/refusal	Removes component 4.

Table M2. The nine experimental conditions and what each isolates.

Conditions 1-3 form the core comparison reported in Section 4.1. Condition 4 is the leakage gate (Section 4.6), which excludes the alternative explanation that the harness merely leaks the answer through

metadata. Conditions 6-9 form the component ablation reported in Section 4.3. The fifth ablation - policy router - was evaluated through a separate routing-disabled run.

3.4 Component ablation design

To attribute the harness effect to its constituent components, we ran a leave-one-out ablation: starting from the full harness, each component was removed in turn while all others were held fixed. The drop in score relative to the full harness is the marginal contribution of the removed component. Ablations were run across all four models on the full 200-item benchmark. Average contributions and per-model contributions are reported in Section 4.3. This design is what permits us to move from showing that the harness works to showing why it works.

3.5 Cross-jurisdiction design

To test whether the findings reflect a property of South Australia rather than a general property of governance harnesses, we repeated the core comparison on a Victorian corpus and benchmark. The system was extended through a `JurisdictionConfig` abstraction: a single configuration object names the seed authorities, the authority allow-list, the governance prompt template, and the escalation rules. No code paths differ between SA and Victoria. Two models (GPT-5.4-mini and Sonnet 4.6) were evaluated on both jurisdictions under the model-only and full-harness conditions.

3.6 Scoring and statistical methods

Each response was scored deterministically by a Python rubric on five dimensions, each worth up to 3 points: (i) must-include terms, the fraction of expected key terms present in the response; (ii) must-not-include terms, with full credit if no forbidden terms appear and zero otherwise; (iii) escalation behaviour, with full credit when observed escalation intent matches the item's expected value; (iv) refusal behaviour, with full credit when observed refusal intent matches the item's expected value; (v) source references, with credit awarded for the presence and precision of citations to the expected authoritative sources. The five sub-scores sum to a maximum of 15 and are linearly rescaled to a 0-100 composite. The rubric is implemented in `src/sa_harness/scoring/scorer.py` and operates entirely on the response text and the benchmark item's known-correct fields. It does not call an evaluation model. Style-based scoring limitations (e.g., whether richer prose changes term-presence detection) are discussed in Section 5.4. Statistical significance for paired comparisons was computed by bootstrap resampling (5,000 resamples, 95% confidence intervals on the mean delta). A delta is reported as significant when its 95% CI excludes zero.

The total cost of the evaluation programme was approximately USD 181 over 21,753 API calls and 57.4M tokens, billed in USD at provider list prices. The reproduction cost for the n=200 core evaluation alone is approximately USD 70.38.

4. Results

4.1 The harness effect is universal and statistically significant

We begin with the core comparison: model-only versus generic RAG versus the full SA harness. Table 1 presents composite scores for all four models, with bootstrap 95% confidence intervals on the harness-minus-model-only delta. Figure 3 visualises the same data.

Model	Model only	Generic RAG	SA harness	Δ (harness-MO)	95% CI	Sig.
Claude Opus 4.7	43.2%	59.5%	65.3%	+22.2pp	[19.2, 25.3]	Yes
Claude Sonnet 4.6	55.6%	66.1%	71.2%	+15.6pp	[12.8, 18.0]	Yes
GPT-5.5	60.8%	70.7%	70.2%	+9.4pp	[6.7, 12.2]	Yes
GPT-5.4-mini	66.9%	69.1%	72.4%	+5.5pp	[3.6, 7.5]	Yes

Table 1. Composite scores by model and condition, with harness deltas and 95% confidence intervals (n=200).

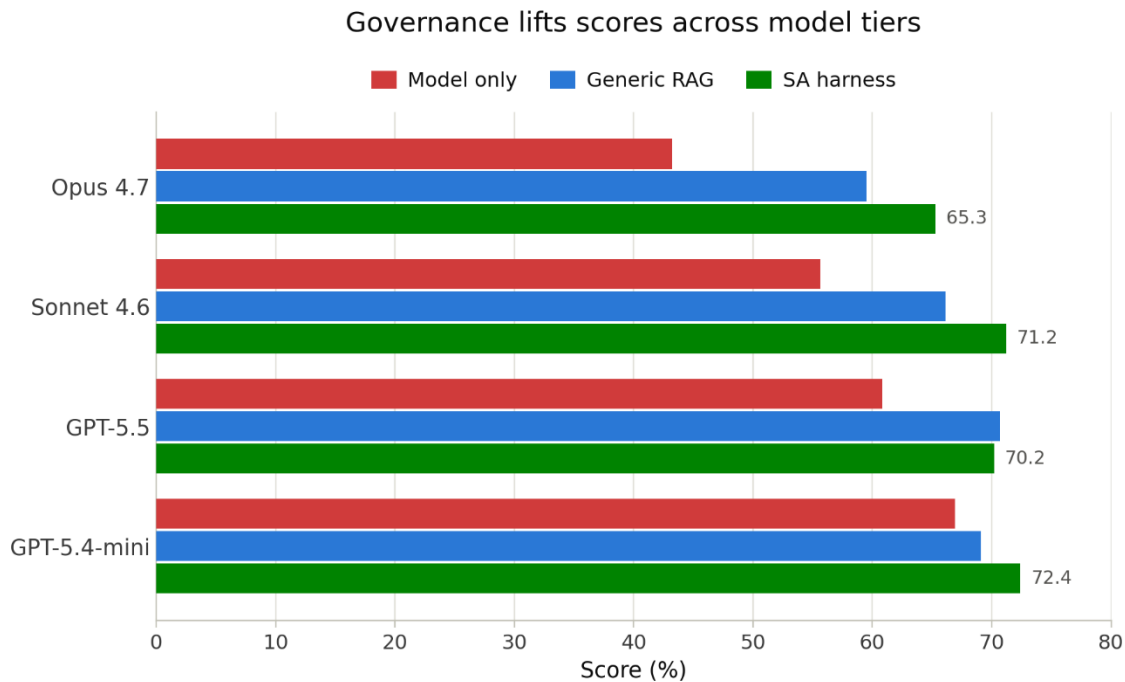


Figure 3. Composite scores by model and condition (n=200). For each model the bars show model-only (baseline capability), generic RAG (retrieval only), and SA harness (full governance layer). Every harness bar exceeds its corresponding RAG bar, and every RAG bar exceeds its corresponding model-only bar.

The key finding from Table 1 and Figure 3 is that the harness effect is not isolated to one model family or one capability tier. It is present in both Anthropic and OpenAI models, across a 23-point range of baseline capability. The smallest delta (GPT-5.4-mini, +5.5pp) is significant at $p < 0.05$ by the bootstrap test. The largest (Opus 4.7, +22.2pp) represents more than half of the model's own baseline score. The harness also

outperforms generic RAG for every model, indicating that the governance functions add value above and beyond retrieval.

4.2 Inverse relationship between baseline score and harness delta

A striking pattern emerges from the deltas in Table 1: they are not constant across models, they shrink as baseline score rises. Opus 4.7, the lowest-scoring baseline at 43.2%, gained 22.2pp. GPT-5.4-mini, the highest-scoring baseline at 66.9%, gained only 5.5pp. Figure 4 plots this relationship.

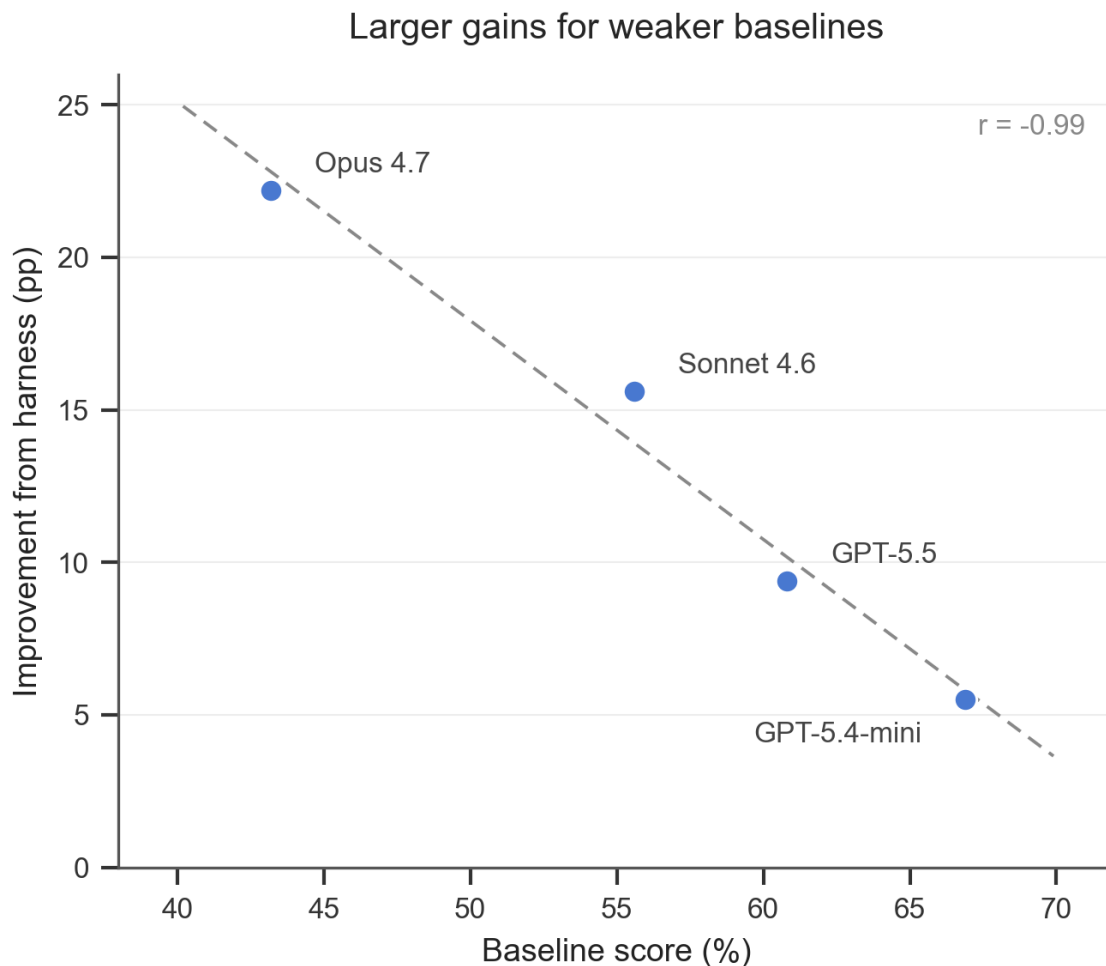


Figure 4. Inverse relationship between baseline score and harness delta. The x-axis is the model-only baseline composite score on this benchmark. The y-axis is the harness delta in percentage points. The four observed models fall almost perfectly on a line with Pearson $r = -0.988$. Lower-scoring baselines on this benchmark gain proportionally more from the harness.

The correlation between baseline score and harness delta is $r = -0.988$. We offer two complementary, partly observationally equivalent interpretations and stop short of choosing between them on the basis of $n=4$ models. The first is a substitution interpretation: lower-scoring baselines lack the implicit governance priors (habits of citation, hedging, and refusal) that higher-scoring baselines have picked up during alignment training, and the harness supplies those priors externally. The second is a ceiling-effect interpretation: the harness raises every model toward a ceiling set largely by the corpus and the rubric, so

models that start closer to that ceiling have less remaining headroom to gain. The data are consistent with both interpretations. With only four models in a narrow score band and a maximum observed harness score of $\sim 72\%$ (well short of the rubric's 100% ceiling), we cannot cleanly separate them. The interpretive caveat matters for the policy reading: if the pattern is wholly a ceiling effect, the practical claim that harness investment compresses cost-quality gaps in this band still holds, but its mechanism is different from a 'governance substitution' story. We return to this in Section 4.5.

4.3 Component mechanism: why the harness works

Establishing that the harness works is one thing. Identifying which of its components carry the value is a different and policy-critical question. If governance value were spread evenly across all five components, the engineering recommendation would be to build all five carefully. If, on the other hand, value were concentrated in one or two components, the engineering recommendation would be to build those components rigorously and treat the remainder as routine. The leave-one-out ablation described in Section 3.4 was designed to answer this question. Table 2 ranks the five governance components by their average marginal contribution across all four models, and reports per-model contributions.

Rank	Component	Avg.	Opus 4.7	Sonnet 4.6	GPT-5.4-mini	GPT-5.5
1	Citation enforcement	+5.9pp	+4.9	+9.7	+5.5	+3.5
2	Source-authority gate	+4.7pp	+3.9	+5.4	+4.5	+4.9
3	Governance system prompt	+2.9pp	+5.8	+4.5	+0.6	+0.4
4	Escalation / refusal	+0.8pp	+0.2	+1.8	-0.3	+1.4
5	Policy router	+0.6pp	+0.0	+0.2	+0.0	+2.1

Table 2. Ranked component contributions to the harness effect, average and per model.

Two findings from Table 2 and Figure 5 carry most of the mechanistic story. First, two components, citation enforcement and the source-authority gate, together contribute an average of +10.6pp, more than two-thirds of the mean harness effect. Citation enforcement is the single largest contributor and is positive for every model. The source-authority gate is the most uniform contributor, positive for every model with the smallest cross-model variance. This pair identifies the mechanism of governance value: it comes from binding outputs to a curated, audited evidence base. Other components matter, but these two do most of the work.

Second, the relative importance of the components varies systematically across models in a way that mirrors the pattern of Section 4.2. For Sonnet 4.6, citation enforcement (+9.7pp) and source authority (+5.4pp) together sum to +15.0pp - roughly 97% of Sonnet's total harness effect of +15.6pp. For Opus 4.7, the same two components sum to +8.8pp - only 39% of Opus's total effect of +22.2pp. Opus's gain is explained much more by the governance prompt (+5.8pp) and to a lesser extent by an interaction term we cannot decompose with this design. We caveat the percentage-of-effect figures throughout this section: the leave-one-out ablation attributes marginal contributions to each component assuming they act roughly

additively. They do not always: for GPT-5.4-mini, the sum of the five individual leave-one-out deltas is approximately +10.3pp against a measured full-harness effect of +5.5pp, i.e., approximately 183% of the total. This indicates negative interactions between components (removing one component may make another work harder), and our design cannot separately identify those interactions. The qualitative ranking (citation enforcement and source authority dominate, the governance prompt matters most for the lowest-scoring baseline, escalation and the policy router are small contributors on this benchmark) holds despite this caveat. Specific percentage-of-effect numbers should be read as indicative rather than precise. The substantive observation stands: stronger-baseline models gain mostly from binding to legitimate evidence. Lower-baseline models additionally need the explicit epistemic posture that the governance prompt supplies. This has implications for how harness investment should be tuned to the model tier in deployment.

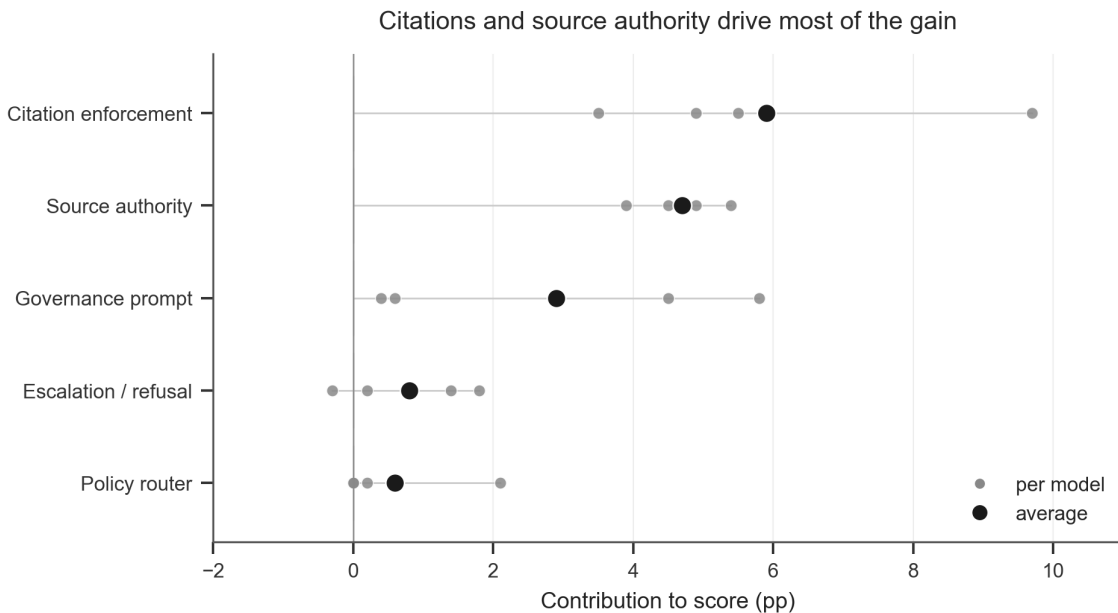


Figure 5. Component contribution to the harness effect, by model. Horizontal bars show the marginal score loss when each component is removed from the full harness. Citation enforcement and the source-authority gate dominate across all four models; the governance prompt is dominant for models with lower baseline scores but negligible for those with higher baseline scores.

The mechanism finding maps directly onto an investment ranking. Citation enforcement and source-authority verification should be built first and engineered most rigorously: they are the two components that pay back across the full capability range, and they are also the two components most amenable to engineering audit. The governance prompt, escalation logic, and policy router matter, but their value is more context-dependent and harder to demonstrate on a general-purpose benchmark. We return to this ranking as an explicit policy recommendation in Section 5.2.

4.4 Cross-jurisdiction transfer

A natural objection at this point is that the findings might reflect specific properties of South Australia (its corpus, its query distribution, its institutional structure) rather than a general property of governance

Sovereign AI Through Governance Harnesses

harnesses applied to public-sector advisory work. To address this, we replicated the core comparison on a Victorian corpus and benchmark, with no code changes, using only the JurisdictionConfig abstraction described in Section 3.5.

Model	SA MO	SA harness	SA Δ [95% CI]	VIC MO	VIC harness	VIC Δ [95% CI]
GPT-5.4-mini	66.9%	72.4%	+5.5pp [3.6, 7.5]	64.1%	71.8%	+7.7pp [4.0, 11.4]
Claude Sonnet 4.6	55.6%	71.2%	+15.6pp [12.8, 18.0]	60.4%	69.7%	+9.3pp [4.2, 14.6]

Table 3. Cross-jurisdiction transfer: model-only and harness scores in South Australia and Victoria.

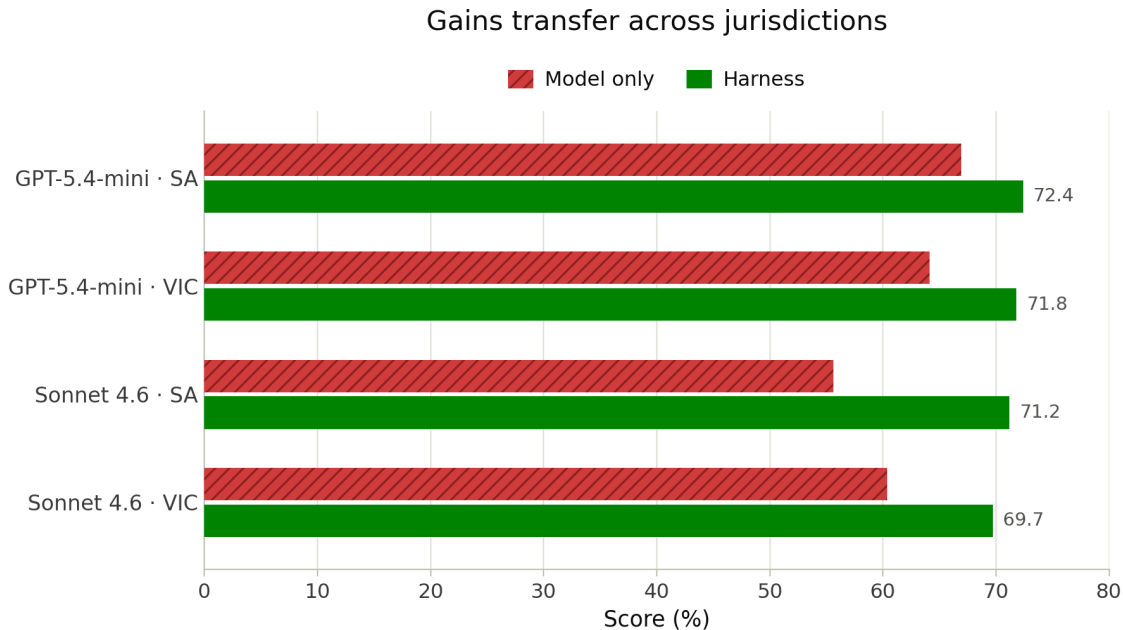


Figure 6. Cross-jurisdiction harness effect for two models. South Australian and Victorian model-only and harness scores are shown as paired bars. Both jurisdictions exhibit statistically significant harness deltas with overlapping effect sizes. Transferring the system required only a new JurisdictionConfig and a fresh corpus.

Table 3 and Figure 6 show that, for both models tested, the harness effect is statistically significant in Victoria (95% CIs exclude zero) and falls within or close to the range of the South Australian effect for the same model. The Victorian effect is somewhat larger for GPT-5.4-mini and somewhat smaller for Sonnet 4.6, but the qualitative finding (significant, substantial, single-digit to double-digit improvement) holds in both jurisdictions. Crucially, the transfer required no engineering changes beyond curating the Victorian corpus and editing a JurisdictionConfig file. The harness pattern therefore appears portable rather than bespoke. The methodology travels.

4.5 Cost and the harness

Section 4.2 implies but does not directly answer a more pointed question: if lower-scoring baselines gain most from the harness, and if cheaper models in this test set happen to start with higher baseline scores,

Sovereign AI Through Governance Harnesses

what does that imply for procurement? Combining the harness scores with provider list prices yields one of the most policy-relevant findings in the study. Table 4 reports approximate cost per benchmark answer (in USD at May 2026 provider list prices) under the harness condition, and Figure 7 places each model on a cost-vs-score plane.

Model	Harness score	Cost / answer (USD)	Relative cost
GPT-5.4-mini	72.4%	~USD 0.05	1x
GPT-5.5	70.2%	~USD 0.31	6x
Claude Sonnet 4.6	71.2%	~USD 0.70	14x
Claude Opus 4.7	65.3%	~USD 1.08	22x

Table 4. Approximate cost per benchmark answer and harness score by model (May 2026 list prices).

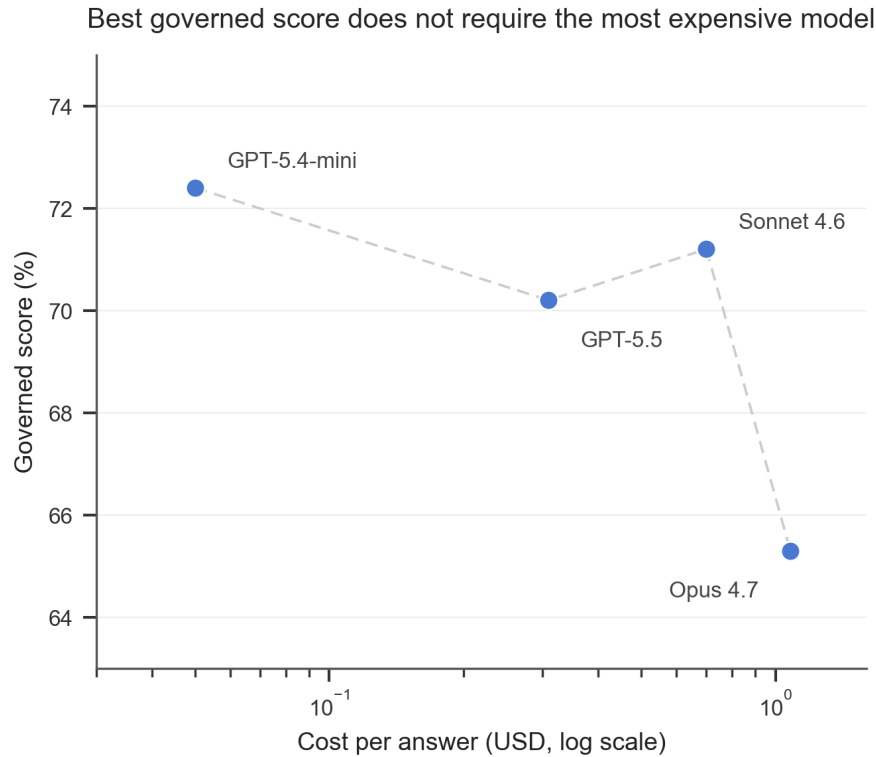


Figure 7. Cost and harness score for the four models tested. Each point is one model. The x-axis is cost per benchmark answer in USD on a logarithmic scale. The y-axis is harness composite score. In this test set the cheapest model produced the highest score (reflecting baseline ordering: GPT-5.4-mini also had the highest baseline) and the most expensive model produced the lowest.

In this test set, model price and harness score were not aligned: the cheapest model (GPT-5.4-mini) produced the highest harness score, and the most expensive (Opus 4.7) produced the lowest. We are cautious about generalising this pattern. Three factors specific to our test set are jointly responsible. First, GPT-5.4-mini happens to begin with the strongest model-only baseline of the four on this advisory-style benchmark, which is a property of how recent training updates land on this content rather than a general claim about model size or price. Second, the harness raises every model toward a ceiling set largely by the corpus and the rubric, compressing differences across models that more capability-demanding

benchmarks would amplify. Third, the cost figures depend on May 2026 provider list prices, which move frequently. The 22x multiplier in particular will shift as providers reprice. The durable qualitative claim is that, in advisory use cases with a curated corpus and a well-engineered harness, model price and model utility decouple sharply enough that a procurement strategy optimised for headline model capability can be strictly dominated by one that pairs an efficient model with a well-engineered harness. The specific finding 'cheapest model wins' is an artefact of this particular four-model panel and the current provider price structure. The durable finding is the decoupling itself.

4.6 Controls: the leakage gate

Before drawing implications we must exclude a methodological alternative: that the harness effect is an artefact of document titles or URLs leaking the correct answer to the model through the prompt. The metadata-minimal condition was designed for this test. It provides only document titles and URLs, with no body text. If metadata leakage were driving the harness effect, the metadata-minimal score should approach the harness score. Table 5 reports the comparison; Figure 8 visualises it.

Model	Metadata-minimal	Generic RAG	Harness	Verdict
Claude Opus 4.7	56.0%	59.5%	65.3%	CLEAR
Claude Sonnet 4.6	59.9%	66.1%	71.2%	CLEAR
GPT-5.4-mini	67.4%	69.1%	72.4%	CLEAR
GPT-5.5	61.9%	70.7%	70.2%	CLEAR

Table 5. Leakage gate: metadata-minimal condition versus generic RAG and the full harness.

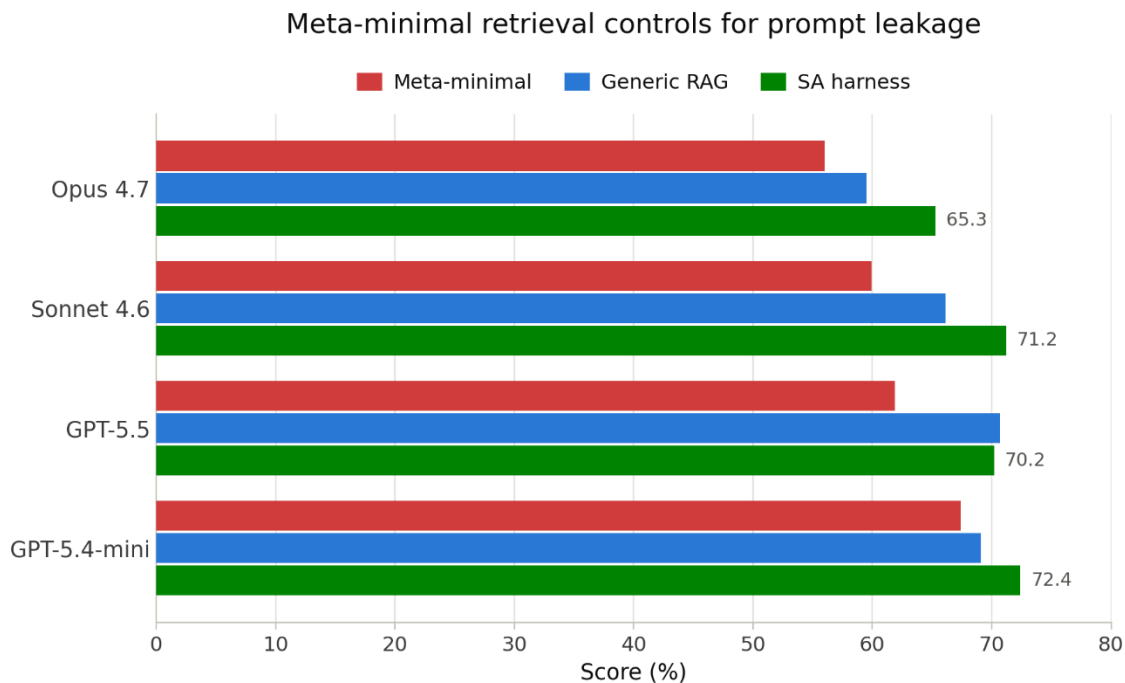


Figure 8. Leakage gate. For every model the metadata-minimal condition (titles and URLs only) scores well below the full harness, supporting the interpretation that the harness effect derives from substantive content delivered through the retrieved sources and the governance contract, not from metadata cues.

For every model, the metadata-minimal score falls clearly below the harness score, with gaps ranging from 5.0pp to 11.3pp. The leakage gate is therefore clear: the harness effect derives from substantive content and the governance contract, not from metadata cues. We can now move to the discussion with confidence that the headline result is not an artefact.

5. Discussion

5.1 The mechanism: governance is a separable layer

The pattern across Sections 4.1-4.3 supports a claim: governance is a separable property, supplied by an external layer rather than emerging from model capability. Three lines of evidence converge on this reading. First, the harness improved every model (across two providers and a 23-point capability range), indicating that no model in our test set supplies the relevant governance properties on its own. Second, the inverse relationship between baseline score and harness delta ($r=-0.988$) is consistent with the externally supplied governance substituting for the implicit governance priors that higher-scoring baselines acquire during alignment training, although as discussed in Section 4.2 a ceiling-effect interpretation is also consistent with the data and the two cannot be cleanly separated from $n=4$ models. Third, the component ablation localises whichever of the two mechanisms is operating: citation enforcement and source authority, the two most structural, evidence-binding components, contribute most uniformly across the score range, while the governance prompt, the most prior-like component, contributes most to the lowest-scoring baseline and least to the highest. Subject to the additivity caveat of Section 4.3, the harness is best read as an external implementation of the governance discipline that better-aligned models partially internalise.

This separability matters for AI sovereignty because it implies that the autonomy property can be built without the capacity property. Sub-national governments cannot build frontier models, but they can build the layer that converts those models into accountable instruments. Critically, the value that this layer adds is concentrated in two engineering-tractable components (citation enforcement and source-authority verification), is portable across jurisdictions (Section 4.4), and arrives at a tiny fraction of the cost of premium model access (Section 4.5).

5.2 Policy implications

The empirical pattern translates into three concrete recommendations for sub-national governments designing sovereign-AI investments for public-sector advisory use.

First, prioritise governance engineering over model procurement. Table 4 shows that, in this test set, paying 22x more for the most expensive model yielded a lower harness score than pairing an efficient model with a well-engineered harness. As Section 4.5 cautions, the specific cost multiplier is sensitive to current provider list prices and to which models happen to anchor the panel. The durable claim is the decoupling of price from harness score, not the precise figure. For the use cases evaluated here, the highest-value investment is in building the harness that wraps the models, not in obtaining the most expensive ones.

Second, within harness investment, prioritise citation enforcement and source authority. For Sonnet 4.6, the two largest components account for 97% of the total harness delta. For GPT-5.5, the figure is 89%, and for Opus 4.7, 39%. For GPT-5.4-mini, the per-component contributions sum to more than 100% of the total harness effect, indicating that the components interact non-additively and that some of the gain is shared between them (Section 4.3). Across the cross-model average they contribute +10.6pp out of the +13.2pp total effect. They are also the two components most amenable to engineering rigour and external audit, because their behaviour is structurally constrained (you can inspect every citation against the corpus) rather than stylistically encouraged. A jurisdiction with constrained engineering resources should fund these first.

Third, treat methodology rather than artefact as the durable investment. The Victorian replication required no engineering changes beyond a new JurisdictionConfig and a curated corpus. Once a jurisdiction has built the harness pattern, applying it to a new domain, agency, or peer jurisdiction is a configuration exercise. This implies that inter-government investment (at the level of state-and-territory collaborations, or sub-national networks such as Regional Australia Institute or the COAG framework) can compound across jurisdictions in a way that model procurement cannot.

We emphasise the limits of these recommendations. They apply to advisory use cases of the kind we evaluated. They do not extend to creative, agentic, multi-turn, or high-autonomy applications. And they presume the existence of a curated jurisdictional corpus of the kind described in Section 3.1. The cost finding in particular depends on current provider list prices, which change frequently, and on the relative position of the four models tested at this point in time. The qualitative finding, that the harness compresses model-quality differences and changes the procurement calculus, is more durable than any specific cost-multiplier figure.

5.3 Relationship to Indigenous data governance

We state a hard boundary. This work is not a substitute for, nor a model of, Indigenous data governance. The harness pattern presented here governs the use of externally sourced AI capacity against state-curated corpora. It does not address ownership of Indigenous data, consent for its use, or community control over downstream applications. The CARE principles for Indigenous data governance (Collective benefit, Authority to control, Responsibility, and Ethics) are the appropriate framework for that domain and require Indigenous-led design from the outset. We note that the harness architecture could, in principle, support Indigenous-led governance (its policy router, authority gate, and escalation logic are general), but only if Indigenous communities lead the design of the corpus, the authority list, and the escalation rules. Nothing in this concept note should be read as proposing otherwise.

5.4 Limitations

- Researcher-authored benchmark items. The 200 SA items and 50 Victorian items were authored by the research team rather than sampled from real public-sector advisory traffic. This may bias the benchmark toward queries the team finds intuitive and away from genuine edge cases.
- Deterministic rubric-based scoring with style sensitivity. All ~12,600 responses were scored by a deterministic Python rubric (Section 3.6) over expected term presence, citation references, and structured escalation/refusal fields. The rubric is reproducible and free of model-jury bias, but it may systematically favour responses whose surface form better matches the expected-term lists,

independent of whether the underlying advisory content is genuinely more or less correct. We have not measured this sensitivity directly.

- No human adjudication. We did not run a human-rated sub-sample to calibrate the rubric against expert judgements of advisory quality. Calibration against human raters on a stratified sub-sample is the most important methodological extension we have not yet done.
- Single-turn only. Every condition was a single prompt and a single response. Multi-turn dynamics (clarification, follow-up, drift) were not evaluated.
- Two jurisdictions, both Australian, with an unbalanced n. South Australia and Victoria share an institutional template, and the Victorian replication used n=50 items against the SA n=200. The Victorian sample is large enough to test transfer but too small to assess all of the Section-4 findings at the resolution achievable in SA. Transfer to genuinely different legal-administrative contexts (US states, EU regions, lower-middle income jurisdictions) is not demonstrated here.
- Closed-provider model panel and small model n. All four evaluated models are closed-weight frontier models from two providers (Anthropic, OpenAI). The four-model panel is too small to fit anything beyond a simple linear trend to the inverse-effect finding, and the cost results depend on which closed-provider models happen to anchor the panel at May 2026 prices. Whether the patterns persist with open-weight models (Llama, Qwen, Mistral) or smaller models below the GPT-5.4-mini tier is an open question that the planned replication described in Section 5.5 is designed to address.
- Incomplete corpus coverage. The SA corpus contains 951 curated sources; the Victorian corpus 913. Both are a substantial fraction of, but not equivalent to, the full jurisdictional information environment. Benchmark items that fall outside the corpus will favour model-only conditions.
- Component independence assumption. The leave-one-out ablation attributes marginal contributions assuming components act roughly additively. They do not always: for GPT-5.4-mini the per-component contributions sum to ~183% of the measured full-harness effect, indicating negative interactions our design cannot separately quantify (Section 4.3). Specific percentage-of-effect figures should be read as indicative.
- Ceiling effects on the inverse pattern. The maximum harness score observed in our data is ~72%, well below the rubric's 100% ceiling. With only four models in a narrow score band, we cannot fully separate a 'governance substitution' interpretation of the inverse pattern from a 'ceiling effect' interpretation (Section 4.2). This affects mechanistic, but not policy, conclusions.
- GPT-5.5 fusion regression. The harness x compressed-retrieval fusion condition produced a small regression on GPT-5.5, indicating a model-specific interaction we do not yet understand and that warrants targeted investigation.
- Cost figures depend on current provider list prices. The 22x cost ratio reflects May 2026 USD list prices and will move as providers reprice. The qualitative pattern (decoupling of model price from harness score) is more durable than the specific multiplier.

5.5 Future work

- Human adjudication on a stratified sub-sample to calibrate the automated scorer.
- Field-sourced benchmark items derived (with appropriate privacy controls) from anonymised public-sector advisory queries.
- Multi-turn evaluation including clarification and follow-up dynamics.

- Replication with open-source models (Llama, Qwen, Mistral families) to test whether the inverse harness effect persists at lower capability tiers and to remove the dependency on closed-provider list prices.
- Transfer beyond Australia: at least one non-Westminster jurisdiction to test whether the JurisdictionConfig abstraction generalises to fundamentally different administrative structures.
- A public-sector pilot deployment, evaluated against service-quality and accountability outcomes rather than benchmark scores alone.

6. Conclusion

This concept note has presented an evaluation methodology and a prototype governance harness for sub-national public-sector AI use, together with empirical results from approximately 12,600 scored responses across four frontier models, nine experimental conditions, two Australian jurisdictions, and five component ablations. The argument proceeds in six steps. Governments are being told to invest in expensive sovereign models. We argue instead that governance is a separable layer that can be built independently of models. The harness improves every model tested and improves the lowest-scoring baselines most, a pattern consistent with both a governance-substitution and a ceiling-effect reading. The mechanism is concentrated in two components, citation enforcement and source-authority verification, which together account for between 39% and 97% of the measured harness effect depending on the model, subject to the additivity caveats discussed in Section 4.3. In our four-model panel, model price and harness score decoupled sharply: the cheapest model produced the highest harness score, at roughly 1/22 the per-answer cost of the most expensive. The result transferred to Victoria (n=50, two models) with zero code changes. The actionable implication is that sub-national governments should invest in citation enforcement and source authority as the first harness investments.

The bottom line for sub-national policymakers is that sovereign AI value, for the public-sector advisory use cases that dominate day-to-day demand, is principally a governance-engineering investment, not a model-procurement decision. A small, auditable harness over commercially available frontier models converts capacity into autonomy. This is a narrower form of sovereignty than the term sometimes implies, but it is one that sub-national governments can actually build, audit, and operate. And on the evidence reported here it is also the form that yields most of the practical value at a small fraction of the cost of the alternative.

References

- Aaronson, S. A. (2025). The dangers of AI nationalism and beggar-thy-neighbour policies. *Journal of World Trade*, 59(3), 397–430.
- Aliberti, M., Cappella, M., & Hrozensky, T. (2019). *Measuring space power: A theoretical and empirical investigation on Europe*. Springer.
- Aliberti, M., Cappelli, O., & Praino, R. (2023). Power, state and space: Conceptualizing, measuring and comparing space actors. *Studies in Space Policy*, Vol. 35. Springer.
- Calzada, I. (2025). Datafied democracies & AI economics unplugged: Technopolitics in smart cities and datafied network states. Springer. <https://doi.org/10.1007/978-3-032-11888-2>
- Cappelli, O. (2010). Pre-modern state-building in post-Soviet Russia. In R. Hill & O. Cappelli (Eds.), *Putin and Putinism*. Routledge.
- Hoeffler, C., & Mérand, F. (2024). Digital sovereignty, economic ideas, and the struggle over the digital markets act: A political-cultural approach. *Journal of European Public Policy*, 31(8), 2121–2146. <https://doi.org/10.1080/13501763.2023.2294144>
- Ishkhanyan, A. (2025). The sovereignty-internationalism paradox in AI governance: Digital federalism and global algorithmic control. *Discover Artificial Intelligence*, 5, Article 123. <https://doi.org/10.1007/s44163-025-00374-x>
- Krasner, S. D. (1999). *Sovereignty: Organized hypocrisy*. Princeton University Press.
- Krishnamurthy, V. (2025). Anchoring digital sovereignty. *Chicago Journal of International Law*, 25(2), 417–477.
- Sartori, G. (1970). Concept misformation in comparative politics. *American Political Science Review*, 64(4), 1033–1053.
- Sheng, L. (2022). *Big Tech firms and international relations: The role of the nation-state in new forms of power*. Springer. <https://doi.org/10.1007/978-981-19-3682-1>
- Srivastava, S. (2017). *Hybrid sovereignty in world politics* [Doctoral dissertation, Northwestern University]. ProQuest Dissertations Publishing.
- Srivastava, S. (2022). Corporate sovereign awakening and the making of modern state sovereignty: New archival evidence from the English East India Company. *International Organization*, 76(3), 690–712. <https://doi.org/10.1017/S002081832200008X>
- Srivastava, S. (2023). Algorithmic governance and the international politics of big tech. *Perspectives on Politics*, 21(3), 989–1000. <https://doi.org/10.1017/S1537592721003145>
- Srivastava, S., & Bullock, J. (2024). AI, global governance, and digital sovereignty (arXiv:2410.17481). arXiv. <https://arxiv.org/abs/2410.17481>
- Veale, M., Matus, K., & Gorwa, R. (2023). AI and global governance: Modalities, rationales, tensions. *Annual Review of Law and Social Science*, 19, 255–275. <https://doi.org/10.1146/annurev-lawsocsci-020223-040749>
- Weymouth, S. (2025). Digital disintegration: Techno-blocs and strategic sovereignty in the AI era. *International Organization*, 79(S1), S57–S70. <https://doi.org/10.1017/S0020818325101070>

Appendix A. Reproducibility and resource footprint

- Total API calls across the evaluation programme: 21,753.
- Total tokens processed: ~57.4 million.
- Estimated total compute spend: ~USD 181 (May 2026 provider list prices).
- Cost to reproduce the n=200 core evaluation only: ~USD 70.38.
- Total scored responses: ~12,600.
- South Australian corpus: 951 curated sources, 200 benchmark items.
- Victorian corpus: 913 curated sources, 50 benchmark items.
- Significance testing: paired bootstrap, 5,000 resamples, 95% confidence intervals.

Appendix B. Architectural notes

The harness is implemented as a thin Python layer with explicit module boundaries for each of the five components described in Section 2.2. Retrieval uses a vector index over the curated corpus. The source-authority gate filters retrieval candidates by domain allow-list prior to ranking. Citation enforcement is implemented as a structured-output contract validated post-generation against the retrieved source set. Escalation and refusal are implemented as deterministic rules over the (domain, risk) pair returned by the policy router. The governance system prompt is loaded from the JurisdictionConfig and prepended to every generation. The total harness implementation is small enough to be audited end-to-end by a single reviewer in a working day.



**Flinders
University**



Jeff Bleich Centre
for Democracy and
Disruptive Technologies

Contact us
jbc@flinders.edu.au