PoliScore:
A Framework for Transparent, Auditable Governance

Draft 0.6 — December 7, 2025

This document is a vision statement for what PoliScore could become: an open, rigorous, and scalable framework for evaluating legislation and legislative behavior. The system described here is not yet fully implemented; the live features on `poliscore.us` *represent early prototypes of a much broader aspiration. Bringing this framework to its full potential will require resources, collaborators, and sustained support from those who believe in the value of transparent, accountable governance.*

Abstract

Artificial intelligence has fundamentally transformed our ability to read, categorize, and reason over large volumes of text. Nowhere is this transformation more consequential than in democratic governance, where legislative text— the primary artifact of policymaking—remains abundant, complex, and largely unanalyzed at scale. Despite its centrality, there exists no widely accepted, non-partisan standard for evaluating the structural quality of legislation or the long-run performance of the elected officials who shape it.

As AI models proliferate across the ideological spectrum, it becomes trivial to measure a legislator's alignment with any worldview; yet it remains exceedingly difficult to construct an AI system that meaningfully justifies corruption or governance failure. With rigorous benchmarks and transparent evaluation criteria, biased or captured models can be quickly exposed. This makes AI not only a tool for interpreting policy, but a structural counterweight to corruption: auditable, stress-testable, and capable of revealing patterns of behavior invisible to human-scale analysis.

PoliScore operationalizes this opportunity. It provides a principled, reproducible framework for evaluating legislative text, grounded in an explicit seven-pillar model of policy quality and a scalable aggregation pipeline for deriving legislator-level impact measures. Its architecture consists of three components: *PoliBench*, an open-source suite for benchmarking an AI model's policy-reasoning capabilities; *PoliScore*, an open-source interpretation engine for fetching, evaluating, and aggregating bill text; and *PoliScore Web*, a closed-source platform that delivers these capabilities to practitioners, researchers, journalists, and the public.

Together, these tools convert legislative text into structured civic knowledge—quantitative scores, qualitative analyses, and audit-ready evaluations—usable by citizens, researchers, nonprofits, journalists, legislative staff, and oversight institutions. By grounding political AI in

transparent methodology and open benchmarks, PoliScore aims to ensure that AI strengthens democratic accountability rather than undermining it.

Contents

1 Introduction

1.1 AI as a Counterweight to Corruption

Artificial intelligence has fundamentally changed what is possible in textual analysis. Tasks that once required teams of researchers—categorizing documents, extracting structure from unstructured text, synthesizing patterns across thousands of pages—are now solvable with commodity hardware and off-the-shelf models. AI's ability to interpret, compare, and aggregate large volumes of text is not a marginal improvement; it is a fundamental transformation. It enables a scale and precision of political analysis that was previously unimaginable.

This matters because legislative text is the beating heart of democratic governance. Every congressional session introduces upward of sixteen thousand bills—most of which are never analyzed in depth by journalists, academics, or the public. These bills collectively contain the intentions, priorities, incentives, and structural choices that shape real human outcomes. They are also a gold mine for data analysis: highly formalized, machine-readable documents that reflect the operational logic of government. AI is uniquely suited to read them all.

And yet, despite the centrality of these texts, public policy still lacks a widely accepted, nonpartisan standard for evaluating the quality of legislation before it is enacted. Public policy shapes the daily lives of individuals and the long-term trajectory of societies, but we possess no unified methodology for assessing the structural soundness of a bill, its expected societal impact, or the cumulative performance of elected officials in terms ordinary voters can understand. The gap between the importance of legislation and our ability to evaluate it is one of the great unaddressed failures of modern governance.

AI enters precisely at this fault line.

As AI models continue to diversify, we will soon have systems representing every major political ideology across the spectrum. This is not speculative; it is an inevitable property of open-source model ecosystems, fine-tuning, and market incentives. Once such models exist, analyzing a legislator's conformance to a particular ideology becomes almost trivial. With minimal tooling, a motivated analyst could run the entire U.S. Congress through an ideological alignment pipeline in a single weekend for well under $100. At scale, such analysis can be performed for pennies.

But this development leads to an unexpected—and profound—conclusion. While it is easy to build an AI model that reflects a particular ideological worldview, it is extremely difficult to construct one that credibly frames legislative corruption as a net positive for society. Even if someone attempted to engineer such a worldview, its internal contradictions would be obvious under scrutiny. With sufficiently rigorous benchmarks and test suites, corrupt or captured models can be exposed quickly and unambiguously. The very act of trying to justify corruption becomes an empirical liability.

This is the paradox: The same technological force that can be used to persuade, manipulate, or obscure the truth can also be used to reveal patterns of corruption that were previously invisible. Because AI models can be audited, stress-tested, compared, and benchmarked, they become natural allies in the fight for transparent governance. Their biases are measurable. Their interpretations are reproducible. Their methods are inspectable.

Therefore, AI—when used to categorize, interpret, and aggregate legislative text—is not merely a tool for understanding politics. It becomes a structural counterweight to corruption itself.

In a political environment where trust is eroding and information is abundant but disorganized, AI provides something rare: a scalable, falsifiable, systematic way to evaluate the gap between what legislators *say* and what they *do*.

1.2 From Vision to System Design

If AI is to serve as a structural counterweight to corruption, then it must operate within a framework that is transparent, principled, and reproducible. The challenge is not merely to read legislative text, but to evaluate it against standards that reflect societal values, institutional safeguards, and long-term public welfare. In other words, AI requires a scaffold: a conceptual architecture that constrains interpretation, enforces rigor, and provides a shared language for evaluation. Without such a scaffold, AI risks drifting toward persuasion, mimicry, or ideological capture.

PoliScore provides that scaffold. It translates the high-level promise of AI-driven governance analysis into a concrete, structured methodology. Where the preceding discussion identifies the *opportunity*—a world where AI can interpret and audit the entirety of legislative activity—PoliScore answers the core design question: *How can we build such a system? And how can we do it in a way which reinforces, rather than deteriorates, our existing democratic institutions?* What makes a bill structurally sound? What constitutes evidence-based governance? How should competing societal impacts be weighed? And how do we distinguish between policy quality, ideological preference, and corruption?

To address these questions, PoliScore introduces a coherent conceptual system grounded in two insights:

1. **Policy evaluation requires explicit criteria.** Without a normative framework, AI becomes a mirror for the data it consumes. PoliScore's seven-pillar model provides the interpretive backbone necessary for consistency, comparability, and transparency.
2. **Legislative performance is an aggregation problem.** Legislators do not exist in isolation from the bills they sponsor, co-sponsor, amend, or vote upon. Their civic impact emerges only when those interactions are systematically aggregated across time. AI makes such aggregation computationally tractable at national scale.

Through this combination of structured evaluation and scalable aggregation, PoliScore transforms the philosophical idea that AI can reveal patterns, incentives, and institutional behavior into a practical analytical system. It gives AI *rules*, *tasks*, and *benchmarks* that tether its outputs to democratic and institutional values rather than partisan narratives. The framework directs the power of large-scale textual analysis toward clarity, accountability, and public understanding rather than persuasion.

In short, PoliScore operationalizes the principle that political AI must be evaluative, audit-ready, and grounded in explicit methodological commitments—not merely generative or performative.

1.3 The Architecture of PoliScore

The PoliScore ecosystem is intentionally structured as an *open core*—a set of public, MIT-licensed tools that enable transparent, auditable analysis of legislative text—paired with a privately maintained application layer that sustains long-term development. This architecture consists of three primary components:

1. **PoliBench** PoliBench is an open-source, MIT-licensed benchmarking suite designed to measure an AI model's capacity for evaluating legislation. It provides structured tasks derived from PoliScore's seven-pillar framework, including assessments of problem clarity, evidence support, governance integrity, systemic risk, and predicted sectoral impacts. By offering reproducible benchmark tasks, PoliBench enables researchers, developers, and the public to evaluate how well different AI systems interpret legislative text and to expose biases, inconsistencies, or ideological capture. In essence, it defines what "good policy reasoning" looks like for AI.
2. **PoliScore** PoliScore is an open-source, MIT-licensed interpretation pipeline for fetching, processing, and evaluating bill text. It operationalizes the seven-pillar policy quality model, generates both quantitative impact scores and qualitative natural-language analyses, and aggregates those bill-level evaluations into legislator-level measures such as Intended Societal Impact scores and parameterized summaries. PoliScore is the analytical engine that turns legislative text into structured, interpretable civic knowledge.
3. **PoliScore Web** PoliScore Web is a closed-source platform built on top of the open core. It provides the user interface, database infrastructure, authentication, operational tooling, and long-term sustainability mechanisms necessary to deliver PoliScore to the public. While the analytical logic remains fully open-source, this layer ensures that the broader ecosystem can be maintained, hosted, and scaled as a durable public resource. PoliScore Web exists to support the mission of the open-source components—not to enclose them.

Together, these components form a coherent architecture: PoliBench defines and tests the capabilities required for trustworthy policy reasoning; PoliScore performs those reasoning tasks at scale; and PoliScore Web delivers those results to practitioners, researchers, journalists, and citizens.

1.4 Who is this For?

The PoliScore ecosystem is designed not merely as a technical exercise, but as an enabling infrastructure for a wide range of civic, academic, journalistic, and institutional users. By converting the full corpus of legislative text into interpretable, auditable metrics, PoliScore provides three broad categories of value:

1. Transparency and Accountability

- **Citizens** gain an evidence-based view of legislative behavior, allowing them to see not just what politicians say, but what they structurally support through their bill interactions.
- **Watchdog groups and anti-corruption organizations** can use PoliScore and PoliBench to detect patterns of governance failure, hidden incentives, or legislative behavior inconsistent with public commitments.
- **Journalists** receive a scalable analytical tool for contextualizing votes, identifying trends across sessions, and grounding political narratives in measurable policy impact.

2. Research and Methodological Rigor

- **Policy researchers and academics** gain a reproducible framework for evaluating policy quality across time, domains, and jurisdictions.
- **AI researchers** gain benchmark tasks and stress tests capable of revealing how well different models reason about governance—a domain traditionally considered too nuanced for automation.
- **Government oversight bodies** can use the system to audit legislative portfolios or evaluate the policy literacy of AI systems used internally.

3. Practical Decision-Making

- **Nonprofits and advocacy organizations** can rapidly assess whether a bill aligns with their mission or introduces systemic risks or governance vulnerabilities.
- **Legislative staff** can use PoliScore to identify similar historical bills, compare policy structures, or anticipate the downstream consequences of proposed amendments.
- **Voters and civic educators** can use the system to explain the real-world impacts of legislation in clear, accessible language.

Across these domains, the central value proposition is consistent: PoliScore transforms legislative text—a dense, technical, often impenetrable artifact—into structured civic knowledge that is scalable, falsifiable, and aligned with the public interest. By grounding AI analysis in explicit methodologies and public benchmarks, the system helps ensure that political AI enhances democratic accountability rather than undermining it.

2 Foundational Principles of Policy Quality

2.1 Human Needs as the Basis of Societal Outcomes

At the core of any political system lies a simple and empirically grounded truth: human beings have universal, predictable needs. These include, at minimum, the requirements for survival (food, water, shelter, health, safety) and the conditions for flourishing (education, opportunity, autonomy, stability, and participation in society).

This principle is not ideological. It traces broadly through the history of philosophy and public policy, from classical notions of basic welfare to modern frameworks such as:

- Maslow's hierarchy of needs (Maslow, 1943),
- Sen's capabilities approach (Sen, 1999),
- Rawlsian justice as fairness (Rawls, 1971),
- Nussbaum's list of central human capabilities (Nussbaum, 2006),
- the UN Human Development Index (UNDP, 1990),
- the OECD Better Life Index (OECD, 2011).

The implication is straightforward: policies ultimately exist to alter societal conditions in ways that affect these human needs. Therefore, any concept of policy quality must begin with the recognition of these universal determinants of human well-being.

2.2 The Purpose of Public Policy: Maximizing Societal Benefit

If human needs are universal, then public policy can be defined as:

> *A structured intervention intended to alter social, economic, or political conditions to maximize societal well-being, minimize harm, and ensure long-term stability.*

This definition integrates insights from multiple traditions:

- Utilitarianism — maximizing aggregate well-being.
- Rawlsian justice — ensuring fairness and protection for the least advantaged.
- Capabilities theory — expanding real freedoms and opportunities.

- Institutional economics — ensuring efficiency, stability, and low transaction costs (North, 1990).
- Governance theory — designing institutions that are transparent, accountable, and resilient (Ostrom, 1990).

From these traditions emerges a balanced, non-ideological goal:

> *Good policy is that which improves human outcomes without creating disproportionate harm, fragility, inequality, or institutional dysfunction.*

This is the foundational goal on which the PoliScore evaluative framework is built.

2.3 From First Principles to a Standard Rubric of Policy Quality

If

(a) humans have predictable needs, and

(b) the purpose of policy is to maximize societal benefit while minimizing harm,

then it follows that policy quality can be systematically evaluated according to how effectively a proposal moves society toward those outcomes.

PoliScore formalizes this into seven universal dimensions derived directly from these foundational principles:

1. **Problem Clarity & Causal Validity**
   Does the policy accurately diagnose the underlying issue and target the relevant causal mechanisms?
2. **Evidence Base & Empirical Support**
   Is the proposed intervention supported by empirical research, historical precedent, or meaningful comparative data?
3. **Implementation Feasibility**
   Can existing institutions realistically execute the policy given resource, logistical, administrative, and temporal constraints?
4. **Economic Efficiency & Fiscal Sustainability**
   Does the policy use resources responsibly, minimize waste, and avoid unsustainable long-term obligations?

5. **Distributional Impact & Fairness**
   How are benefits and burdens distributed across populations, and does the policy unjustifiably disadvantage certain groups?

6. **Governance Integrity & Institutional Risk**
   Does the policy maintain transparency, accountability, and resilience while minimizing opportunities for corruption or abuse?

7. **Unintended Consequences & Systemic Risk**
   Does the policy introduce fragility, perverse incentives, or cascading failures that undermine the intended outcomes?

These seven pillars are not ideological criteria. They are derived from the intersection of:

- moral philosophy,
- economics,
- development theory,
- organizational behavior,
- risk analysis,
- governance studies,
- institutional design.

Taken together, they form a non-partisan, foundational theory of policy quality designed explicitly for computational and rubric-based evaluation.

2.4 Why This Framework Is Necessary

Existing institutions (such as the CBO, independent economic modeling groups, think tanks, and academic departments) evaluate policy through isolated lenses:

- cost impacts,
- economic forecasts,
- ideological alignment,
- advocacy positions,
- program evaluation after implementation.

None provide a unified, interdisciplinary, pre-implementation standard for determining whether legislation is well-designed, feasible, fair, evidence-based, and structurally beneficial.

PoliScore seeks to fill this gap by offering:

- a consistent methodology,
- grounded in decades of cross-disciplinary research,
- applicable directly to bill text,
- measurable, reproducible, and benchmarkable,
- independent of ideology.

This framework is the intellectual foundation for the PoliScore grading system and the PoliBench benchmark suite.

2.5 A New Field: Policy Quality Engineering

By synthesizing philosophical foundations, institutional economics, governance theory, and modern AI evaluation, PoliScore introduces what is effectively a new discipline:

> **Policy Quality Engineering:** the systematic, reproducible evaluation of public policy according to universal human needs, societal benefit, and institutional feasibility.

This document is ultimately an invitation to help build that field. The rest of the paper develops both the theoretical and practical aspects of this discipline and outlines how it could be extended from bill-level analysis to legislator-level performance grading, given sufficient technical and institutional support.

3 The PoliScore Framework

The PoliScore framework provides a systematic, non-partisan method for evaluating the quality of public policy based on its expected real-world impact, feasibility, and alignment with universal human needs. Derived from the foundational principles articulated in Section 2, the framework operationalizes these concepts through seven core evaluation dimensions, each representing a necessary component of high-quality legislation.

Each dimension is designed to be:

- philosophically grounded,

- empirically motivated,
- institutionally relevant,
- computationally assessable, and
- applicable directly to legislative text.

Together, these dimensions form a comprehensive rubric for determining whether a policy is constructed in a way that maximizes societal benefit while minimizing harm, inefficiency, and unintended consequences.

3.1 Problem Clarity & Causal Validity

Effective policy begins with a clear, accurate understanding of the problem it seeks to address. Vague or misdiagnosed problems lead to interventions that fail to produce meaningful improvements or that target symptoms rather than causes.

A policy demonstrates clarity and causal validity when:

- the problem is explicitly defined and empirically measurable;
- the underlying causal mechanisms are identified;
- the proposed intervention plausibly affects those mechanisms;
- the theory of change is coherent and logically sound.

Policies that rest on untested assumptions, moral panic, or ideological narratives—rather than a valid causal model— score poorly on this dimension.

3.2 Evidence Base & Empirical Support

High-quality policy proposals demonstrate clear grounding in empirical research, comparative case studies, or validated theoretical frameworks. This dimension evaluates whether the intervention is supported by evidence that similar approaches have succeeded elsewhere, or whether its expected outcomes are consistent with existing knowledge.

Relevant criteria include:

- citation of empirical findings or documented precedents;
- alignment with established best practices in relevant fields;

- avoidance of claims contradicted by available data;
- transparency about uncertainty and knowledge gaps.

Policies lacking empirical grounding may rely on wishful thinking or unproven assumptions, increasing the risk of unintended harm.

3.3 Implementation Feasibility

Even a theoretically sound policy can fail if it is infeasible to implement. Feasibility depends on the capacity of institutions, agencies, and local systems to carry out the policy's mandates with available resources, logistics, workforce, technology, and time.

This dimension evaluates:

- administrative complexity and burden;
- clarity of agency responsibilities;
- resource requirements (financial, human, technical);
- timeline realism;
- reliance on unavailable or overextended infrastructure;
- potential bottlenecks or bureaucratic overload.

Policies that impose unrealistic workloads, require nonexistent infrastructure, or centralize responsibility in ways that exceed institutional capacity receive low scores.

3.4 Economic Efficiency & Fiscal Sustainability

Public policy must allocate resources responsibly, avoid generating structural inefficiencies, and maintain long-term fiscal viability. This dimension evaluates whether the benefits of the policy justify its costs, whether incentives are aligned with real-world behaviors, and whether it avoids unnecessary waste or economic distortion.

Considerations include:

- long-term funding stability;
- cost-benefit alignment;
- administrative overhead;

- market distortions or inefficiencies;
- externalities (positive or negative);
- dynamic economic effects and sustainability over time.

Policies that rely on implausible revenue assumptions, produce excessive deadweight loss, or generate persistent deficits score poorly.

3.5 Distributional Impact & Fairness

Public policies distribute benefits and burdens across different groups within society. A high-quality policy should not impose disproportionate harm on specific populations or create unjustifiable imbalances in who gains and who loses. This dimension evaluates how a policy's effects are spread across income levels, regions, industries, and demographic groups, and whether these effects align with broadly accepted principles of fairness and responsible governance.

Key considerations include:

- which groups receive the primary benefits of the policy;
- which groups bear the costs or risks;
- whether the distribution of impacts is reasonable and transparent;
- whether the policy inadvertently worsens existing disadvantages;
- whether the policy shifts burdens onto populations with limited capacity to absorb them.

This pillar does not require or assume equal outcomes. Instead, it assesses whether the distribution of impacts is justified, defensible, and consistent with the stated goals of the policy, and whether any imbalances introduce meaningful risk or undue harm.

3.6 Governance Integrity & Institutional Risk

Policies exist within complex governance structures. This dimension evaluates whether a proposal strengthens or undermines institutional integrity, transparency, accountability, and the rule of law.

Key criteria:

- clarity of authority and decision-making processes;
- adequacy of oversight and accountability mechanisms;

- risks of corruption, abuse of power, or regulatory capture;
- concentration of unregulated authority;
- resilience to political manipulation;
- clarity in compliance requirements.

Policies that concentrate discretionary power in unaccountable agencies, lack oversight, or create opportunities for corruption receive lower scores.

3.7 Unintended Consequences & Systemic Risk

Complex systems often respond unpredictably to policy interventions. This dimension assesses the extent to which a policy may produce harmful unintended consequences, including perverse incentives, moral hazard, market failures, bureaucratic overload, or cascading systemic risks.

Evaluative criteria include:

- creation of fragile dependencies;
- incentive misalignment;
- spillovers into adjacent systems;
- risk of black markets or evasion;
- increased systemic fragility or bottlenecks;
- insufficient fail-safes or fallback mechanisms.

Policies that appear beneficial in theory but introduce hidden structural costs or vulnerabilities receive lower scores.

3.8 Sectoral Impact Model and Overall Impact to Society

The seven pillars described above characterize *structural policy quality*. To connect these abstractions to concrete societal outcomes, PoliScore introduces a sectoral impact model and a scalar "Overall Impact to Society" score for each bill.

For each bill b, PoliScore defines a set of sectoral impact scores:

$$I_{b,s} \in [-100, 100] \cup \{\text{N/A}\}$$

for each sector s in a fixed set $\mathcal{S}$ that includes, at minimum:

- Agriculture and Food
- Education
- Transportation
- Economics and Commerce
- Foreign Relations
- Government Efficiency and Management
- Healthcare
- Housing
- Energy
- Technology
- Immigration
- National Defense
- Crime and Law Enforcement
- Wildlife and Forest Management
- Public Lands and Natural Resources
- Environmental Management and Climate Change

Each sector score is rated from -100 (very harmful) to 0 (neutral) to $+100$ (very helpful), or marked as N/A when the sector is not meaningfully affected by the bill.

From this sectoral vector, PoliScore defines an "Overall Impact to Society" score:

$$I_b^* \in [-100, 100],$$

representing a synthetic estimate of the bill's aggregate impact on societal well-being, considering the interaction of all affected sectors.

In practice, $I_{b,s}$ and I_b^* are produced by an evaluator (currently an AI model in prototype experiments) that:

1. reads the bill text,
2. applies the seven-pillar framework to understand structure, feasibility, and risks,
3. performs constrained research where appropriate, and

4. assigns sectoral scores and an overall score with an accompanying natural-language justification.

The sectoral model and overall score serve as the bridge between abstract policy quality and concrete societal impact, and they are the primary inputs into legislator-level aggregation described in Section 4.

Summary of the Framework

Together, these seven structural dimensions and the sectoral / overall impact model form a holistic, first-principles representation of policy quality. They address not only a policy's intent and potential benefits, but also its feasibility, fairness, evidence base, institutional risks, and long-term sustainability.

The PoliScore framework is designed to:

- provide structured, rational evaluation of legislation;
- enable reproducible scoring across policies and time;
- support AI-assisted and human-driven legislative analysis;
- inform policymakers, researchers, and the public;
- reduce reliance on ideological or partisan heuristics.

This framework also provides the theoretical foundation for both the bill-to-legislator pipeline (Section 4) and the PoliBench benchmark suite (Section 5), which tests whether AI systems can reliably interpret and evaluate policy quality along these dimensions.

Definition (Intended Societal Impact). *Intended Societal Impact (ISI)* is the aggregate, expected effect on society implied by the set of bills a legislator sponsors, co-sponsors, or votes for or against, weighted by interaction type. ISI represents the legislator's revealed policy intent under the PoliScore evaluation framework: the world their legislative record points toward, assuming their positions were enacted.

4 From Bills to Legislators: The PoliScore Impact Pipeline

PoliScore is ultimately designed to answer two questions:

1. **Bill-level question:** What is the predicted impact of this bill on society?

2. **Legislator-level question:** Over the course of their recent history, what has this legislator actually done for (or to) the public through their legislative actions?

This section describes the multi-stage pipeline that connects the seven-pillar framework and sectoral impact model to legislator performance scores and letter grades. As of today, parts of this pipeline exist in prototype form; the full design is presented here as a target architecture for a more mature, well-resourced system.

4.1 High-Level Pipeline Overview

Conceptually, the end-to-end pipeline can be summarized as:

1. Bill text is segmented and analyzed under the PoliScore framework.
2. A sectoral impact vector and an "Overall Impact to Society" score are produced for each bill.
3. Legislator–bill interactions (sponsorship, co-sponsorship, and votes) are collected.
4. Each legislator's Intended Societal Impact score is computed as a weighted function of the overall bill impact scores.
5. Legislators and bills receive letter grades derived from their scalar impact scores.
6. Parameterized natural-language summaries translate these metrics into human-readable narratives.

Figure 1 depicts this process.

Figure 1: High-level PoliScore pipeline from bill text to legislator grades.

Throughout this paper, a legislator's score is framed primarily in terms of **Intended Societal Impact**. This quantity represents the aggregate, expected effect on society implied by the legislator's pattern of sponsorships, co-sponsorships, and votes, assuming their preferred positions on bills were consistently enacted. It therefore reflects the legislator's *revealed policy intent* and the structural quality of the policies they support, rather than the empirical outcomes of laws that were actually implemented. In Section 4.5 we introduce complementary metrics—such as realized impact and pillar- or sector-level performance—that provide additional perspectives on legislative behavior.

While this pipeline is currently prototyped using AI systems as the primary evaluators, each step is defined in a way that could be executed by trained human analysts following the same rubric, or scaled up by future AI systems that have been explicitly validated on PoliBench.

4.2 Bill-Level Scoring Prompt and Outputs

In its early instantiation, PoliScore uses a structured evaluation template for bill-level scoring. The evaluator (AI model or human) is asked to behave as a non-partisan oversight committee, reading full bill text and producing:

- sectoral impact scores $I_{b,s}$,
- an "Overall Impact to Society" score I_b^*,
- short and long-form explanatory reports, and
- a self-rated confidence score.

The template for the Stats section is:

```
Stats:
Score the following bill on the estimated impact to the United
States upon the following criteria, rated from -100 (very harmful)
to 0 (neutral) to +100 (very helpful) or N/A if it is not relevant.

Agriculture and Food:  <score or N/A>
Education:  <score or N/A>
Transportation:  <score or N/A>
Economics and Commerce:  <score or N/A>
Foreign relations:  <score or N/A>
Government Efficiency and Management:  <score or N/A>
Healthcare:  <score or N/A>
Housing:  <score or N/A>
Energy:  <score or N/A>
Technology:  <score or N/A>
Immigration:  <score or N/A>
National Defense:  <score or N/A>
Crime and Law Enforcement:  <score or N/A>
Wildlife and Forest Management:  <score or N/A>
Public Lands and Natural Resources:  <score or N/A>
Environmental Management and climate change:  <score or N/A>
Overall Impact to society:  <score or N/A>
```

Additional sections provide a descriptive title, a short report, a long report referencing specific bill provisions, and a numeric confidence score. In a fully realized implementation, these outputs would be logged, versioned, and made publicly inspectable.

4.3 Aggregating Sectoral Scores into Overall Impact

The "Overall Impact to Society" score I_b^* may be produced in one of two ways:

1. **Direct overall assessment.** The evaluator directly assigns I_b^* based on holistic reasoning over the bill and its context.
2. **Derived aggregation.** The evaluator provides only sectoral scores $I_{b,s}$, and the system computes:
$$I_b^* = f(\{I_{b,s}\}_{s \in \mathcal{S}}),$$
where f is an aggregation function (e.g., a weighted average of non-N/A sector scores).

In practice, PoliScore can use both approaches simultaneously: the evaluator provides a direct overall score, and the system cross-checks it against a derived aggregate. Discrepancies can be logged for analysis, calibration, or human review.

A simple, transparent aggregation rule is:

$$I_b^* = \frac{\sum_{s \in \mathcal{S}_b} w_s I_{b,s}}{\sum_{s \in \mathcal{S}_b} |w_s|},$$

where $\mathcal{S}_b$ is the set of sectors marked as relevant (non-N/A) for bill b, and w_s are sector weights. In the earliest implementation, $w_s = 1$ for all sectors, yielding an unweighted average. More sophisticated variants may:

- estimate w_s from public opinion and expert surveys;
- tune w_s based on empirical outcomes of historical legislation;
- incorporate robustness constraints (e.g., penalize extreme scores confined to a single sector).

The key requirement is that the mapping from sectoral scores to overall impact is formally defined, transparent, and stable over the evaluation period.

4.4 Legislator–Bill Interaction Model

Legislators interact with bills in recognizable ways: they sponsor them, co-sponsor them, and vote for or against them. Let:

- $\mathcal{B}$ denote the set of bills,
- $\mathcal{L}$ denote the set of legislators,

- I_b^* denote the overall impact score of bill b,
- $T \in \{\text{Sponsor}, \text{CoSponsor}, \text{VotedFor}, \text{VotedAgainst}\}$ denote the type of interaction.

PoliScore defines weights w_T for each interaction type, reflecting both normative judgment and intuitive voter expectations. A simple weighting scheme currently used in practice is:

$$\begin{aligned} w_{\text{Sponsor}} &= 1.0, \\ w_{\text{CoSponsor}} &= 0.7, \\ w_{\text{VotedFor}} &= 0.5, \\ w_{\text{VotedAgainst}} &= -0.5. \end{aligned}$$

The negative weight for VotedAgainst indicates that opposing a harmful bill ($I_b^* < 0$) generates positive credit, while opposing a beneficial bill ($I_b^* > 0$) generates negative credit.

For a given legislator $\ell \in \mathcal{L}$, let $\mathcal{I}_\ell$ denote the set of all (bill, interaction-type) pairs in which they participate. The legislator's Intended Societal Impact score, ISI_ℓ, is defined as:

$$\text{ISI}_\ell = \frac{\sum_{(b,T)\in\mathcal{I}_\ell} w_T \, I_b^*}{\sum_{(b,T)\in\mathcal{I}_\ell} |w_T|}.$$

This formulation:

- credits legislators for sponsoring and supporting beneficial bills,
- debits them for supporting harmful bills,
- rewards them for opposing harmful bills,
- penalizes them for opposing beneficial bills,
- normalizes scores so that legislators with many interactions are comparable to those with fewer.

Intended Societal Impact should be interpreted as the *expected societal impact of the legislator's policy record if their preferred positions were enacted.* It reflects intent, judgment, and policy quality alignment under the PoliScore framework.

4.5 Alternative and Complementary Legislator Metrics

The aggregation rule above defines a legislator's primary score as a function of bill-level impact values. In PoliScore, this quantity is called the legislator's **Intended Societal Impact** (ISI). While ISI is the central metric, the same pipeline naturally supports additional views of legislative performance.

This subsection introduces three complementary perspectives built on the same primitives: realized societal impact, pillar-level performance, and sector-level performance.

4.5.1 Realized Societal Impact

While ISI measures the consequences of a legislator's *intent*, a natural complement is to measure their contribution to laws that actually entered into force. Let $E_b \in \{0, 1\}$ indicate whether bill b was enacted. A realized variant is then:

$$\text{RealizedImpact}_\ell = \frac{\sum_{(b,T)\in\mathcal{I}_\ell} E_b \, w_T \, I_b^*}{\sum_{(b,T)\in\mathcal{I}_\ell} E_b \, |w_T|}.$$

Where ISI captures "the world they are trying to create," the realized score captures "the world their successful actions have contributed to." The two together provide a fuller behavioral portrait.

4.5.2 Pillar-Level Performance Profiles

Each bill b is evaluated along seven structural pillars of policy quality. Let $D_{b,k}$ denote the score of bill b on pillar $k \in \{1, \dots, 7\}$. A legislator-level pillar profile is defined by:

$$Q_{\ell,k} = \frac{\sum_{(b,T)\in\mathcal{I}_\ell} w_T \, D_{b,k}}{\sum_{(b,T)\in\mathcal{I}_\ell} |w_T|}.$$

The vector $(Q_{\ell,1}, \dots, Q_{\ell,7})$ summarizes how closely a legislator's supported policies align with high-quality design principles: clarity of problem definition, evidence base, implementation feasibility, fairness, governance integrity, and systemic risk. These profiles can be surfaced directly to users (e.g., as tables or radar plots) and incorporated into summary narratives.

4.5.3 Sectoral Performance Profiles

Analogously, for each bill PoliScore defines sectoral impacts $I_{b,s}$ for sectors $s \in \mathcal{S}$. Aggregating these yields:

$$R_{\ell,s} = \frac{\sum_{(b,T)\in\mathcal{I}_\ell} w_T \, I_{b,s}}{\sum_{(b,T)\in\mathcal{I}_\ell} |w_T|}.$$

The vector $(R_{\ell,s})_{s\in\mathcal{S}}$ reveals where a legislator is most helpful or harmful across domains such as healthcare, energy, national defense, housing, or the environment.

Together, the Intended Societal Impact score, its realized-impact counterpart, and these pillar- and sector-level profiles form a multi-dimensional, transparent representation of legislative performance grounded in the PoliScore policy quality framework.

4.6 Letter Grades for Bills and Legislators

To make the results accessible to voters, PoliScore converts raw impact scores into letter grades.

For bills, the letter grade is derived directly from I_b^*; for legislators, from ISI_ℓ or from an alternative legislator-level scalar such as $\mathrm{RealizedImpact}_\ell$. A simple and currently deployed mapping is:

$$\begin{aligned}
\text{A} &: I \geq 40, \\
\text{B} &: 30 \leq I < 40, \\
\text{C} &: 15 \leq I < 30, \\
\text{D} &: 0 \leq I < 15, \\
\text{F} &: I < 0,
\end{aligned}$$

where I is either a bill's overall impact score I_b^* or a legislator's aggregate score.

These thresholds are intentionally simple and interpretable. They are not meant to be perfect statistical constructs, but to provide a consistent, monotonic mapping between scalar impact scores and an intuitive grading scale familiar to the public.

4.7 Parameterized Legislator Summaries

Numerical scores and letter grades provide clarity and comparability, but they do not by themselves tell a story. To bridge this gap, PoliScore uses parameterized natural-language prompts that consume:

- a legislator's aggregate impact statistics by sector,
- their overall letter grade,
- a curated list of their most consequential bill interactions.

A generalized template is:

```
The United States {{politicianType}} {{fullName}} has been evaluated
based on recent legislative performance and has received the
following policy area grades (scores range from -100 to 100):

{{stats}}

Based on these scores, this legislator has received the overall
letter grade: {{letterGrade}}. You will be given bill interaction
summaries of this politician's recent legislative history,
sorted by their impact to the relevant policy area grades.
Please generate a layman's, concise, three paragraph, {{analysisType}},
highlighting any {{behavior}}, identifying trends, referencing
specific bill titles (in quotes), and pointing out major focuses
and priorities of the legislator. Focus on the policy areas
with the largest score magnitudes (either positive or negative).
Do not include the legislator's policy area grade scores and
do not mention their letter grade in your summary.

{{billInteractions}}
```

Where:

- `politicianType` is “Senator” or “House Representative”,
- `fullName` is the legislator’s name,
- `stats` is a textual rendering of sectoral scores,
- `letterGrade` is the overall grade,
- `analysisType` is one of “endorsement”, “mixed analysis”, or “harsh critique”, chosen based on the letter grade,
- `behavior` describes whether to emphasize accomplishments, alarming behavior, or both,
- `billInteractions` is a list of the legislator’s most impactful bill interactions by sector and magnitude.

This design allows the system to vary tone in a controlled, transparent way while keeping the underlying scoring rules fixed. In a fully realized implementation, these summaries would be clearly labeled as model-generated analysis grounded in the underlying metrics, not as neutral journalism.

4.8 AI Behavior, Training Data, and Non-Partisan Prompts

PoliScore's bill-level evaluations are intended to be non-partisan and rooted in "Overall Impact to Society". Three properties of modern language models make this goal plausible when combined with careful prompt design:

1. **Frequency weighting.** Model predictions are influenced by patterns frequently observed in training data. Because training corpora contain large volumes of books, encyclopedias, and mainstream reporting, many of the model's "default" positions track the most commonly articulated views in those sources.

2. **Contextual weighting.** Given a specific prompt, the model preferentially draws from portions of its training data that match the requested context (e.g., technical analysis vs. jokes). When prompted as a non-partisan oversight committee asked to cite scientific studies, official reports, and expert opinions, the model tends to prioritize more authoritative and technical contexts.

3. **Coherence pressure.** To answer PoliScore-style prompts, a model must synthesize scattered facts and arguments into a single coherent narrative that satisfies multiple constraints (sectoral scoring, justification, trade-off analysis). This encourages structured reasoning rather than surface-level slogan repetition.

PoliScore reinforces non-partisanship through explicit design choices:

- prompts emphasize "Overall Impact to Society" rather than partisan advantage,
- the aggregation logic is mechanically neutral and publicly documented,
- the code is released under a permissive open-source license, inviting scrutiny and forks,
- funding and governance of the project aim to remain independent of party structures,
- outputs are transparent and inspectable both at bill and legislator levels.

At the same time, AI systems can inherit biases from their training data, including policy preferences that align with majority public or scientific opinion on issues such as renewable energy, reproductive rights, or gun policy. Rather than denying this, PoliScore presents its outputs as a structured reflection of how a capable model, prompted to be non-partisan, reconciles expert and public consensus with the specifics of legislative text. Users are encouraged to examine the underlying justifications and decide for themselves whether they agree with the resulting grades.

5 The PoliBench Benchmark Suite

polibench

The PoliBench Benchmark Suite is a standardized set of tests designed to evaluate whether AI systems can accurately assess public policy along the seven dimensions of the PoliScore Framework and reason coherently about sectoral and overall societal impacts. Whereas PoliScore provides the conceptual model and pipeline for policy quality and legislator performance, PoliBench operationalizes those ideas into concrete, reproducible tasks that measure an AI system's ability to interpret legislative intent, feasibility, consequences, and institutional design.

PoliBench is not intended as a performance leaderboard for general AI capabilities. Rather, it is a domain-specific benchmark focused on policy reasoning, causal inference, and institutional awareness—areas where existing language models often demonstrate gaps despite strong natural language proficiency. The benchmark enables systematic comparison across AI systems and provides an empirical foundation for evaluating progress in computational policy analysis.

5.1 Motivation

Despite rapid advances in large language models (LLMs), there is currently no standardized method for testing their ability to interpret legislation or assess public policy quality. Existing AI benchmarks measure skills such as:

- question answering (SQuAD, Natural Questions),
- general knowledge (MMLU),
- reasoning (GSM8K, ARC),
- truthfulness (TruthfulQA),
- code generation (HumanEval).

None of these capture the skills required for policy evaluation, such as:

- recognizing ambiguous or misleading problem statements;
- detecting infeasible mandates;
- identifying governance risks;
- reasoning about distributional impacts and sectoral tradeoffs;
- understanding institutional constraints;

- anticipating unintended consequences;
- evaluating evidence claims in context.

Public policy is a systems problem involving economics, governance, logistics, human behavior, and institutional dynamics. PoliBench fills a critical gap by testing whether AI systems can navigate these complexities in a disciplined and consistent way.

5.2 Objectives

PoliBench is designed to achieve five key objectives:

(1) **Evaluate policy-specific reasoning**
Measure whether an AI system can analyze legislative text in ways aligned with the PoliScore pillars and sectoral impact model.

(2) **Provide reproducible, standardized tests**
Ensure that models are evaluated under identical conditions, enabling meaningful comparisons.

(3) **Identify structural weaknesses in AI policy analysis**
Pinpoint which dimensions (e.g., feasibility, unintended consequences) pose the greatest difficulty for current models.

(4) **Support model improvement**
Provide researchers with targeted diagnostics for training and fine-tuning AI systems on legislative reasoning tasks.

(5) **Promote transparency and accountability**
Allow policymakers, academics, and the public to understand the strengths and limitations of AI in this domain.

5.3 Benchmark Structure

PoliBench is organized into seven test suites, one for each structural dimension of the PoliScore Framework, plus optional suites for sectoral impact prediction and legislator-level aggregation consistency. Each suite contains multiple task types, designed to assess both conceptual understanding and applied reasoning.

Core task families include:

- problem clarity and causal validity,

- evidence base and empirical support,
- implementation feasibility,
- economic efficiency and fiscal sustainability,
- distributional impact and fairness,
- governance integrity and institutional risk,
- unintended consequences and systemic risk,
- sectoral and overall impact estimation (optional),
- bill-to-legislator aggregation reasoning (optional).

Each suite includes constructed counterexamples, fictional bills, paired policy comparisons, and scenario-based prompts that allow objective scoring.

5.4 Benchmark Format and Example Tasks

Each PoliBench task is designed to be:

- *deterministic* — clear pass/fail or graded criteria,
- *grounded* — tied to a specific quality dimension,
- *text-based* — directly applicable to legislative language,
- *model-agnostic* — usable by any AI system,
- *transparent* — accompanied by human-annotated rationale and expected answer patterns.

Tasks follow formats such as:

- multiple-choice reasoning tests,
- short-form explanation tasks,
- bill snippet analysis,
- policy comparison tasks,
- error detection tasks (spot-the-flaw),
- numeric prediction of sectoral impacts for simplified scenarios.

Illustrative examples include:

Feasibility Example.
Prompt: A bill requires every rural county (population $< 5{,}000$) to operate a full-service emergency hospital within one year.
Question: Identify the primary implementation challenge.
Expected: Workforce shortages; infrastructure constraints; unrealistic timeline.

Governance Risk Example.
Prompt: A bill grants an agency director unilateral authority to allocate funds "as they see fit," without reporting requirements.
Expected: Lack of oversight; corruption risk; unclear accountability.

Distributional Impact Example.
Prompt: A tax credit is available only to households with mortgage interest payments.
Expected: Benefits flow primarily to homeowners; renters receive no support; regressive impact.

5.5 Scoring and Evaluation

Each model evaluated on PoliBench would receive:

- per-suite scores (0–100),
- a composite PoliBench score,
- diagnostic reports identifying which dimensions require improvement.

Scores reflect:

- correctness,
- reasoning quality,
- internal consistency,
- robustness across variations,
- sensitivity to subtle policy flaws.

By aggregating results across suites, PoliBench would reveal the policy reasoning profile of an AI model and help determine whether it is suitable for deployment within the PoliScore evaluation pipeline.

6 Methodology

The methodology underlying PoliScore is designed to translate legislative text into a structured evaluation across the seven dimensions of policy quality, sectoral and overall societal impact, and finally legislator performance. The process combines computational analysis, domain-specific prompt engineering, and rubric-based scoring to produce a transparent, interpretable assessment of both policies and legislators.

Crucially, the methodology is *process-first*: AI models are current instantiations of the evaluator, but the pipeline itself is defined in a model-agnostic way that could be executed by trained human raters or alternative AI systems. At present, many of these steps exist as prototypes or design documents rather than production-grade components; they are described here to make the blueprint concrete and critiqueable.

The methodological approach consists of five integrated components:

1. Text Preparation,
2. Dimension-Level Policy Evaluation,
3. Sectoral and Overall Impact Scoring,
4. Legislator Aggregation and Grading,
5. Interpretability and Justification.

6.1 Text Preparation

Before analysis, the legislative proposal undergoes standardized preprocessing to ensure that the evaluator receives input in a structured and interpretable form.

6.1.1 Document Segmentation

Legislation is segmented into logical units, such as:

- sections and subsections,
- enumerated provisions,
- definitions,
- mandates,
- authorizations,

- appropriations.

Segmentation allows for focused analysis of complex bills and supports cross-referencing within explanations.

6.1.2 Contextual Metadata

Where available, contextual information is included:

- policy domain (healthcare, tax policy, infrastructure, etc.),
- jurisdiction and legislative session,
- sponsoring entity,
- historical or comparative precedents,
- relevant statutory references.

Metadata helps evaluators identify feasibility constraints, institutional conflicts, and relevant comparators.

6.1.3 Legislative Normalization

Formatting inconsistencies, redundant boilerplate, and non-substantive artifacts (e.g., page headers, XML tags) are removed to reduce noise. The goal is to present the evaluator with a clean representation of the substantive legal content.

6.2 Versioned Legislative Text and Semantic Diffing

Real-world legislation is rarely static. Bills are frequently amended in committee, modified on the floor, reconciled between chambers, or replaced entirely via "strike-and-insert" procedures. Treating a bill as a single, fixed document therefore obscures an important part of the lawmaking process: *how the text evolved over time and who supported which changes*.

PoliScore can be extended to operate over *versioned* legislative text rather than only the final enrolled version. Given a sequence of bill versions $\{b^{(1)}, b^{(2)}, \ldots, b^{(T)}\}$, the system can:

- compute traditional text diffs to identify added, removed, or modified provisions;

- apply *semantic* diffing, using an evaluator to summarize the substantive effect of those changes (e.g., “narrows eligibility to these populations,” “removes the sunset clause,” “adds an enforcement mechanism”);
- update pillar scores and sectoral impacts incrementally as the bill evolves, rather than re-scoring from scratch;
- attribute amendments or version shifts to specific sponsors or coalitions, where metadata is available.

This opens the door to a richer class of analyses. Instead of asking only, “What is the Overall Impact to Society of the final bill text?” PoliScore can also ask, “How did the bill’s Intended Societal Impact change over time, and which actors consistently pushed it toward more or less beneficial outcomes?”

In principle, the same framework used for bill-level scoring can be applied to pairs of versions $(b^{(t)}, b^{(t+1)})$, with the evaluator focused specifically on the marginal change in structure, feasibility, fairness, or systemic risk. This enables semantic diffing that is far more informative than line-based text comparisons, and it aligns naturally with the legislator-level aggregation described in Section 4: amendments and versioned edits become additional interaction events that shape a legislator’s profile of Intended Societal Impact.

6.3 Dimension-Level Policy Evaluation

For each of the seven pillars of the PoliScore Framework, a structured evaluation prompt is applied. Each prompt is designed to assess a specific dimension using criteria derived directly from the theoretical foundations in Sections 2 and 3.

Each dimension evaluation includes:

- **targeted questions** probing specific failure modes,
- **rubric-aligned checklists** (e.g., for feasibility or governance risks),
- **requests for explicit trade-off analysis** where relevant.

The evaluator produces:

- a short narrative assessment,
- a 0–100 numeric score for the dimension,
- optional flags indicating unusual uncertainty or potential contradictions with other dimensions.

Cross-pillar consistency checks (e.g., between feasibility, economic sustainability, and systemic risk) are applied to detect internal conflicts.

6.4 Sectoral and Overall Impact Scoring

Dimension-level evaluations are conceptually upstream of sectoral scoring: understanding problem clarity, evidence base, and unintended consequences informs a more realistic estimate of sector-level impacts.

The sectoral scoring step uses the Stats template and sectoral set $\mathcal{S}$ described in Section 3.8. The evaluator:

1. identifies which sectors are meaningfully affected by the bill,
2. assigns scores $I_{b,s} \in [-100, 100]$ or N/A for each sector,
3. produces an initial overall impact score I_b^*,
4. explains major positive and negative contributions with references to specific provisions.

An aggregation function f is then applied as a cross-check or as a primary method if direct overall scoring is not used. Discrepancies between direct and derived overall scores may trigger:

- automatic confidence reduction,
- a request for reevaluation,
- human review for high-stakes bills.

6.5 Legislator Aggregation and Grading

Once bill-level overall impact scores I_b^* are computed, legislator performance can be evaluated using the interaction model defined in Section 4. The steps are:

1. Construct the set of legislator–bill interactions from roll-call data and bill metadata (sponsorship and co-sponsorship).
2. Apply the weighting scheme w_T to each interaction type T.
3. For each legislator ℓ, compute the Intended Societal Impact score ISI_ℓ as a normalized weighted sum of I_b^*.
4. Optionally compute complementary metrics such as $\mathrm{RealizedImpact}_\ell$, pillar-level profiles $Q_{\ell,k}$, and sectoral profiles $R_{\ell,s}$.
5. Map scalar legislator scores to letter grades via the standard threshold function.

6. Generate sectoral breakdowns of the legislator's performance (e.g., average impact in healthcare vs. energy).
7. Feed these statistics into the parameterized summary prompt to produce a narrative overview.

Because all steps are explicit and documented, users can:

- inspect which bills contributed most to a legislator's score,
- understand how interaction types were weighted,
- recompute scores under alternative weighting schemes if desired.

6.6 Open-Source Implementation and International Adaptation

Although the conceptual framework of PoliScore is jurisdiction-agnostic, any concrete implementation must contend with the specific institutional details of a given legal system: chamber structures, bill numbering schemes, committee processes, roll-call formats, and sector taxonomies. To make adaptation feasible across countries, PoliScore is being developed and released under the permissive MIT license. This choice is deliberate: it allows governments, civil society organizations, researchers, and independent developers to freely reuse, modify, and redistribute the codebase, including in proprietary or jurisdiction-specific deployments, while preserving attribution.

In practical terms, an internationally extensible PoliScore implementation must:

- separate core evaluation logic (the seven pillars, sectoral aggregation, legislator interaction model, and scoring rules) from jurisdiction-specific adapters for:
  - ingesting legislative text and metadata,
  - mapping local committee and chamber structures,
  - interpreting session calendars and bill lifecycles,
  - normalizing roll-call data and sponsor/co-sponsor records;
- expose clear interfaces for customizing:
  - sector lists and weights,
  - default interaction weights for sponsorship and votes,
  - naming conventions and role labels for legislators;
- provide comprehensive documentation and worked examples showing how to stand up a new instance for a different parliament, congress, or assembly.

Because the MIT license imposes minimal restrictions, institutions in other rule-of-law democracies can choose to either extend the mainline project (contributing back improvements and new adapters) or maintain their own branch tailored to local norms and legal structures. In both cases, the underlying policy quality definitions and aggregation mathematics can remain shared, enabling cross-national comparison of bills and legislative behavior while respecting jurisdictional differences in constitutional design and political practice.

6.7 Interpretability and Justification

PoliScore emphasizes transparency and interpretability to ensure users understand *why* a policy or legislator received a given score.

6.7.1 Bill-Level Explanations

For each bill, PoliScore aims to produce:

- a short report summarizing goals and expected impacts,
- a longer, lay-accessible report referencing concrete sections,
- a per-dimension breakdown of strengths and weaknesses,
- a justification for each sectoral and overall impact score.

6.7.2 Legislator-Level Explanations

For each legislator, PoliScore would provide:

- the Intended Societal Impact score and letter grade,
- sectoral performance statistics,
- a list of most influential positive and negative bill interactions,
- a three-paragraph narrative summary generated from the parameterized prompt.

These outputs are intended to empower voters to ask informed questions, not to dictate specific political conclusions.

6.7.3 Consistency and Robustness Checks

To improve trustworthiness, PoliScore incorporates:

- redundant prompts phrased in different ways to test stability,
- spot checks of bill and legislator scores under minor variations,
- adversarial tasks drawn from PoliBench,
- manual review and correction for anomalous outputs.

Summary

The PoliScore methodology provides a rigorous, structured, and interpretable approach to evaluating both policy quality and legislative performance. By combining legislative segmentation, dimension-specific evaluation, sectoral and overall impact scoring, and explicit aggregation rules, PoliScore offers a transparent and reproducible system for assessing the strengths and weaknesses of both bills and the legislators who interact with them. Realizing this methodology at scale will require sustained engineering work, data infrastructure, and institutional stewardship that go far beyond the current unfunded prototype.

7 Comparison to Existing Institutions

Public policy evaluation is not a new endeavor. Governments, academic centers, think tanks, and international organizations have long attempted to analyze the consequences of legislation. However, their approaches are fragmented, domain-specific, and often limited to economic forecasting, post-hoc program evaluation, or ideologically motivated analysis.

This section compares PoliScore to institutions most commonly involved in policy assessment, clarifying how the envisioned framework would complement existing tools while filling an unmet need in pre-implementation policy quality evaluation and systematic legislator performance grading.

7.1 Congressional Budget Office (CBO)

The Congressional Budget Office provides cost estimates and economic projections for federal legislation. Its analyses are widely respected and intentionally nonpartisan. However, by statutory mandate, the CBO:

- does not evaluate whether a policy is fair, feasible, or well-designed;

- does not assess governance risks or institutional fragility;
- does not analyze unintended consequences outside fiscal or macroeconomic domains;
- does not provide judgments about whether a policy is "good" or "poor";
- focuses almost exclusively on budgetary impacts, not policy quality or legislator performance.

How PoliScore differs. In its envisioned form, PoliScore would:

- evaluate seven dimensions of policy quality rather than a single economic dimension;
- focus on pre-implementation design soundness and sectoral societal impacts;
- assess governance structure, feasibility, fairness, and systemic risk;
- aggregate bill impacts into legislator performance scores using transparent rules.

CBO answers: *"What will this cost?"*
PoliScore aims to answer: *"Is this policy structurally sound and societally beneficial, and how does a legislator's record aggregate across such policies?"*

7.2 Think Tanks

Think tanks (e.g., Brookings, Heritage, AEI, CAP, Cato) play a major role in shaping public discourse about policy. They often produce:

- whitepapers,
- policy briefs,
- economic analyses,
- advocacy reports,
- commentary.

However, nearly all think tanks produce work shaped by underlying ideological commitments or donor priorities. As a result:

- evaluations differ radically across institutions;
- frameworks for analysis vary widely;

- there is no unified or nonpartisan definition of policy quality;
- methodology is often opaque or narrative-driven;
- conclusions may be advocacy-oriented rather than diagnostic.

How PoliScore differs. PoliScore:

- does not advocate for policy positions;
- uses a transparent, standardized framework and aggregation model;
- applies the same evaluative criteria to all legislation and legislators;
- grounds analysis in political philosophy, institutional economics, and governance theory rather than ideology;
- produces structured, replicable outputs, not narrative persuasion.

Think tanks answer: *"Is this policy aligned with our values or goals?"*
PoliScore aims to answer: *"How structurally sound is this policy, and what is the aggregate Intended Societal Impact of a legislator's actions?"*

7.3 Academic Public Policy and Political Science Programs

Academic programs teach:

- cost-benefit analysis,
- program evaluation,
- ethics and justice,
- public administration,
- implementation theory.

However:

- methods vary by institution and instructor;
- frameworks are conceptual, not standardized;
- academics rarely evaluate policy before implementation at scale;
- analyses are usually qualitative, not rubric-based and automated;
- no unifying cross-disciplinary "policy quality standard" or legislator aggregation model exists.

How PoliScore differs. PoliScore:

- operationalizes academic theory into a single coherent rubric;
- formalizes seven dimensions into measurable criteria;
- evaluates policy pre-implementation and at scale;
- integrates insights from economics, philosophy, governance, and systems engineering into a concrete scoring pipeline.

Academia answers: *"How should we think about policy and governance?"*
PoliScore aims to answer: *"How well-constructed is this specific policy, and what does a legislator's record look like when evaluated under those standards?"*

7.4 International Organizations and Independent Models

International organizations and independent economic modeling groups evaluate:

- development programs,
- governance indicators,
- effectiveness of existing policies,
- macroeconomic consequences.

Their analyses are valuable but limited in scope. They mainly:

- evaluate implemented programs, not proposed legislation;
- rely on national data and long-term outcomes;
- focus on specific domains;
- analyze macro-level indicators rather than the structure of individual bills or the cumulative record of individual legislators.

PoliScore, if fully developed, would complement these efforts by:

- evaluating legislation prospectively, before implementation;
- operating at the level of bill text and individual legislators;
- providing a cross-domain framework that can be applied uniformly;
- focusing on structural design quality and predicted societal impact rather than only observed outcomes.

7.5 Summary of Differences

Table 1 summarizes the relationships between existing institutions and PoliScore.

Institution Type	What They Evaluate	What They Do Not Evaluate	PoliScore's Intended Contribution
CBO	Budgetary/fiscal impacts	Governance, feasibility, fairness, systemic risk, individual records	Provides holistic pre-implementation structural and impact evaluation, plus legislator aggregation via Intended Societal Impact.
Think Tanks	Ideology-driven arguments	Standardized, nonpartisan evaluation	Supplies neutral, structured scoring across seven dimensions and a transparent bill-to-legislator model.
Academia	Theory, ethics, evaluation methods	Unified applied rubric, automated scoring, legislator grades	Operationalizes theory into a consistent, scalable framework.
International Orgs & Macro Models	Retrospective outcomes; macro projections	Pre-implementation structural analysis of bills; individual legislator histories	Evaluates proposals before enactment and aggregates impacts to individual legislators.

Table 1: Comparison of existing institutions and the envisioned role of PoliScore.

8 Limitations, Risks, and Ethical Considerations

While PoliScore and the PoliBench Benchmark Suite provide a structured and theoretically grounded framework for evaluating public policy and legislative performance, they also introduce methodological, computational, and ethical challenges. Recognizing these limitations is essential for responsible use and for ensuring that PoliScore complements, rather than replaces, democratic decision-making and expert judgment.

In the current, unfunded phase of the project, many of these limitations are particularly acute: there is not yet a large team to manage peer review, adversarial testing, or institutional partnerships. This section therefore serves both as a warning label and as a roadmap for the governance structures that would be needed if PoliScore were adopted more broadly.

8.1 Limitations of the Framework

8.1.1 Incomplete Representations of Policy Context

Legislative text does not always contain:

- administrative history,
- political constraints,
- agency capabilities,
- cultural factors,
- stakeholder incentives,
- implementation environment.

PoliScore evaluates text as written and explicit interaction data, not the full political or institutional context. Real-world outcomes may differ from what the text and modeled impacts suggest.

8.1.2 Dependence on Model Interpretation

PoliScore currently relies on large language models for bill-level evaluation. While PoliBench is intended to validate baseline competence, no model is infallible. AI systems may:

- misinterpret ambiguous sections,
- overlook subtle governance issues,
- fail to identify complex incentive structures,
- inconsistently explain their reasoning,
- exhibit sensitivity to prompt phrasing.

Human oversight remains essential, particularly for high-impact legislation.

8.1.3 Normative Judgments in Pillars and Aggregation

Although the seven dimensions and aggregation rules are derived from widely accepted principles, any evaluative framework contains implicit assumptions. For example:

- principles of "fairness" rely on particular philosophical traditions;
- feasibility depends on assumptions about institutional capacity;
- sector weights and interaction weights embody normative views about what matters most.

PoliScore mitigates this by making all assumptions explicit and configurable, but it is not value-free.

8.1.4 Challenges in Quantifying Qualitative Constructs

Some aspects of policy quality—such as institutional trust, political legitimacy, or cultural acceptance—are inherently difficult to quantify. PoliScore focuses on the structural and impact dimensions that can be reasoned about from text and well-specified criteria, but cannot capture all nuances of real-world politics.

8.1.5 Early-Stage Field

Policy quality engineering is a new field. As such:

- the framework will evolve,
- new dimensions or sub-dimensions may be added,
- weighting schemes may be refined,
- benchmark tasks will need continuous updates.

Versioning and transparent changelogs are therefore essential, especially as more collaborators become involved.

8.2 Risks Associated With AI-Assisted Policy and Legislator Evaluation

8.2.1 Overreliance on AI Outputs

AI-generated evaluations, if misunderstood as authoritative, may:

- overshadow legitimate political debate;
- reduce the perceived role of experts and stakeholders;
- disincentivize democratic deliberation;
- be mistaken for objective truth rather than structured analysis.

PoliScore is designed as an analytical tool, not a source of political mandates.

8.2.2 Risk of Misuse or Politicization

Any scoring system can be misused or selectively weaponized. Risks include:

- cherry-picking PoliScore results to support partisan narratives;
- misrepresenting composite scores without context;
- selectively highlighting favorable dimensions or interactions;
- using grades to target political opponents without acknowledging framework assumptions.

Mitigation requires full publication of methods, dimension-level breakdowns, and underlying bill-level scores, and ideally some form of independent oversight or advisory board.

8.2.3 Model Bias and Training Data Influence

Even when the framework is neutral, AI models may incorporate biases from:

- training data composition,
- institutional assumptions embedded in public discourse,
- coverage biases in scientific literature or media.

PoliScore's focus on "Overall Impact to Society" and evidence-based reasoning pushes models toward mainstream scientific and public consensus on many topics, but this may not align with all ideological perspectives.

8.2.4 Vulnerability to Adversarial Design

Sophisticated actors could attempt to influence model outputs by:

- strategically crafting bill text to appear more beneficial under known evaluation criteria;
- embedding misleading language that exploits LLM weaknesses;
- gaming sector-specific scoring patterns.

Ongoing research, adversarial testing, and human review are required to identify and mitigate such strategies.

8.3 Ethical Considerations

8.3.1 Transparency and Explainability

Users must understand:

- how PoliScore generates evaluations,
- what each dimension and sector score means,
- how the aggregation into legislator grades is performed.

PoliScore therefore emphasizes:

- open access to scoring rules and prompts,
- clear documentation of pipeline steps,
- availability of the underlying bill-level and interaction-level data used for aggregation.

8.3.2 Human Oversight and Democratic Authority

AI models and algorithmic scoring frameworks must not:

- replace elected representatives;
- override democratic decision-making;
- be treated as infallible arbiters of political truth.

PoliScore is intended to help voters, researchers, and policymakers reason more clearly about policy and legislative performance, not to dictate outcomes.

8.3.3 Inclusivity in Framework Development

As PoliScore evolves, its legitimacy depends on:

- peer review,
- collaboration with academics and practitioners,
- multidisciplinary input from economics, law, public health, and other domains,
- feedback from civic organizations and the broader public.

An inclusive development process helps reduce blind spots and increases trust. At present, such a process does not yet exist around PoliScore; building one will require support from institutions that care about non-partisan democratic infrastructure.

Summary

PoliScore introduces a rigorous, first-principles approach to evaluating policy quality and legislative performance, but it is not a substitute for human judgment or democratic deliberation. Recognizing its limitations and potential risks is essential to its responsible use. By emphasizing transparency, interpretability, cross-checking, and academic grounding, PoliScore aims to serve as a constructive analytical tool rather than an authoritative arbiter of political outcomes. To make that aspiration real, the project will need governance, funding, and stewardship that extend beyond its current, single-author prototype status.

9 Integrating Web Search: Benefits, Complexities, and Concerns

As AI systems become increasingly capable of retrieving and synthesizing information from the web, responsible integration of external data into policy evaluation becomes both an opportunity and a challenge. Web search can significantly strengthen PoliScore's analytical depth, especially for feasibility and evidence-based assessment, but it also introduces complexities related to reliability, bias, reproducibility, and governance.

The considerations outlined in this section largely mirror those that apply to AI-assisted analysis in general, but with additional emphasis on preserving the non-partisan and reproducible nature of PoliScore. In the current phase of the project, these ideas are primarily design goals for future, well-resourced iterations.

9.1 Benefits of Web Search Integration

Web search provides access to real-world information that extends beyond the content of legislative text. When properly constrained, it can:

- supply up-to-date empirical data (e.g., workforce statistics, infrastructure capacity),
- clarify agency structures and institutional mandates,
- retrieve historical examples of similar policies,
- contextualize budget numbers and program scales.

These capabilities directly support:

- Pillar 2 (Evidence Base & Empirical Support),
- Pillar 3 (Implementation Feasibility),
- Pillar 4 (Economic Efficiency & Fiscal Sustainability),
- Pillar 6 (Governance Integrity & Institutional Risk).

Web search also helps reduce hallucination by anchoring model outputs in verifiable sources.

9.2 Complexities and Risks

However, integrating web search raises several challenges:

- **Non-reproducibility.** Search results change over time and may vary by geography or search engine configuration.
- **Source quality.** The web contains a mix of official data, academic work, advocacy, and misinformation.
- **Ranking bias.** Search engine algorithms may implicitly prioritize certain narratives or domains.
- **Domain imbalance.** Some policy areas are richly documented, others are sparse.

Without strict guardrails, web search could erode the non-partisan foundation of PoliScore and undermine trust.

9.3 Safeguards and Design Principles

To mitigate these risks, any web-integrated variant of PoliScore should:

- restrict retrieval to vetted source categories (e.g., official government domains, major statistical agencies, peer-reviewed journals, reputable international organizations),
- emphasize retrieval of *facts* (e.g., numeric data, definitions) rather than opinion or advocacy,
- cache retrieved documents and associate them with timestamps and citations for reproducibility,
- expose retrieved sources in public reports so that users can independently evaluate them,
- flag evaluations that rely heavily on sparse or low-confidence external data.

PoliBench can be extended with retrieval-dependent tasks to test model behavior under these constraints.

10 Conclusion and Future Directions

Public policy and legislative performance together define a large part of the lived experience of citizens. Yet until now, there has been no unified, non-partisan, and methodologically rigorous framework for evaluating the structural quality of legislation and aggregating its expected impacts into clear, interpretable assessments of what legislators have actually done.

PoliScore addresses this gap by:

- articulating a principled, interdisciplinary theory of policy quality grounded in human needs, political philosophy, institutional economics, governance theory, and systems thinking;
- formalizing this theory into a seven-pillar evaluation framework and a sectoral impact model that yield an "Overall Impact to Society" score for each bill;
- defining a transparent aggregation pipeline from bill-level impacts to legislator-level Intended Societal Impact scores and letter grades;
- providing parameterized natural-language summaries that help voters understand complex legislative histories;
- introducing PoliBench, a benchmark suite for measuring AI competence in policy reasoning and ensuring a baseline level of evaluator reliability.

Together, these components represent early steps in what may become a broader field: *policy quality engineering*— a discipline focused on the structural, empirical, and institutional soundness of public policy design and on principled aggregation of those designs into measures of legislative performance.

10.1 Opportunities for Future Work

Several avenues for future research and development are evident:

- **Empirical validation.** Applying PoliScore retrospectively to historical legislation and comparing predictions to observed outcomes can help calibrate weights, refine dimensions, and validate predictive power.
- **Collaboration with academia.** Partnerships with universities and research centers can stress-test the framework, refine benchmarks, and extend the theoretical foundations.
- **Sector-specific extensions.** Domain-specific sub-frameworks (e.g., for healthcare, climate policy, tax reform) can be developed atop the core pillars.
- **Richer retrieval and evidence systems.** Carefully constrained and documented use of web-based and curated data sources can strengthen evidence-based evaluation.
- **Internationalization.** Adapting the framework for use in other jurisdictions, legal traditions, and languages can broaden its applicability. Because the PoliScore implementation is released under the MIT license, such adaptations can be carried out by independent institutions that fork or extend the core codebase, reusing the shared definitions of policy quality while substituting local data feeds, institutional mappings, and sector weights. Over time, a constellation of jurisdiction-specific PoliScore deployments could emerge, each tuned to its own rule-of-law environment but grounded in a common evaluative vocabulary.
- **Civic integration.** Tools built on PoliScore could support voters, journalists, advocacy groups, and even legislators themselves in understanding and improving legislative performance.

10.2 A Call for Collaboration and Support

At the time of this writing, PoliScore is not backed by a foundation, a university, or a government grant. It is a person-scale project trying to sketch what a more sane, quality-focused information layer over democratic institutions might look like. As such, there are hard limits to how far it can go without help.

The vision outlined here—a world where voters can see through partisan narratives to the structural quality and predicted impact of the laws being proposed in their name—will require:

- **Engineers and data practitioners** to build and harden the pipelines, data models, and public interfaces.
- **Researchers and academics** to critique the framework, refine the pillars, and help construct PoliBench as a serious domain benchmark.
- **Civic technologists and journalists** to explore how PoliScore-style analysis can be integrated into reporting, voter guides, and accountability tools.
- **Institutional partners**—nonprofits, universities, or public-interest labs—willing to host and steward an open, non-partisan infrastructure.
- **Funders** who recognize that democratic health depends not just on voting procedures, but on the quality of information citizens have about the laws being written and the people writing them.

Because the codebase is MIT-licensed, participation can take many forms: contributing directly, forking and adapting, building independent tools that plug into the same conceptual framework, or simply subjecting the ideas to serious, public critique. The key is that the work moves us away from a politics dominated by memes and outrage, and toward a shared concern with whether our laws are actually well-designed for human flourishing.

10.3 Closing Remarks

The challenges facing modern societies—economic transformation, climate change, technological disruption, public health, and geopolitical instability—demand a new level of clarity and rigor in how we evaluate both policies and the people who make them. PoliScore offers a concrete, transparent, and extensible starting point: a way to move discussions from partisan narratives toward structured analysis rooted in human welfare and institutional robustness.

By framing both bills and legislators in terms of predicted societal impact, and by making every step of the pipeline explicit and inspectable, PoliScore aims to give voters, researchers, and policymakers a clearer view of what our laws are likely to do—and what our representatives have actually done. For now, it is a vision sustained by a single developer. Whether it becomes a durable part of our democratic infrastructure will depend on whether others see value in that vision and decide to help build it.

References

[1] John Rawls, *A Theory of Justice*. Harvard University Press, 1971.

[2] Amartya Sen, *Development as Freedom*. Oxford University Press, 1999.

[3] Elinor Ostrom, *Governing the Commons*. Cambridge University Press, 1990.