PoliScore: From Policy Quality Engineering to Legislative Performance Grading

Version 0.2 — December 2, 2025

Disclaimer: The methodologies, frameworks, and evaluation systems described in this white paper represent an ongoing research and engineering effort. Certain components are partially implemented or in active development. Some production features currently deployed on `poliscore.us` reflect earlier or simplified versions of this framework and should be interpreted as experimental demonstrations rather than complete implementations of the system described herein.

Abstract

This white paper introduces PoliScore, a first-principles framework and computational pipeline for evaluating both the structural quality of public policy and the aggregate legislative performance of elected officials. At the bill level, PoliScore defines a seven-pillar theory of policy quality—problem clarity, evidence support, implementation feasibility, economic sustainability, distributional impact, governance integrity, and systemic risk—derived from foundational insights in political philosophy, welfare economics, institutional theory, and systems engineering. These dimensions are then mapped into a sectoral impact model that scores legislation on its predicted effects across major policy domains (such as healthcare, education, energy, and national defense) and an overall "Overall Impact to Society" metric.

At the legislator level, PoliScore aggregates bill-level impact scores using a transparent model of legislative interactions (sponsorship, co-sponsorship, and roll-call votes), producing scalar performance scores and interpretable letter grades for each legislator. The same framework supports parameterized natural-language summaries that explain a legislator's record in terms accessible to non-expert voters.

PoliBench, a companion benchmark suite, operationalizes the seven-pillar framework into reproducible evaluation tasks that test an AI system's ability to carry out policy reasoning, foresee unintended consequences, identify governance vulnerabilities, and reason about distributional and sectoral impacts directly from legislative text. Although PoliScore is currently implemented using large language models, the underlying methodology is *model-agnostic*: in principle, the same process could be executed by teams of human analysts following a shared rubric or by alternative AI systems.

Together, PoliScore and PoliBench provide the foundation for a new discipline of *policy quality engineering*, while also demonstrating a concrete, end-to-end pipeline from bill analysis to legislator grading. The paper concludes with limitations, ethical considerations, discussion of web-integrated evidence retrieval, and directions for future research and institutional collaboration.

Contents

1 Introduction

Public policy shapes the daily lives of individuals and the long-term trajectory of societies. Yet despite its immense influence, there is no widely accepted, non-partisan standard for evaluating the quality of legislation before it is enacted, nor a unified methodology for summarizing the long-run legislative performance of elected officials in terms ordinary voters can understand.

Voters face an increasingly impossible informational task. Organized misinformation campaigns, partisan media ecosystems, and attention-maximizing social platforms collectively produce a high-noise environment in which the underlying policy record of legislators is difficult to discern. It is easy to surface a few symbolic votes or speeches; it is far harder to understand the cumulative impact of years of legislative activity on real human outcomes.

At the same time, advances in artificial intelligence have made it possible to read and reason over large volumes of legislative text, related reports, and contextual evidence. Used carefully, AI can help bridge the gap between raw policy artifacts and structured, interpretable assessments. Used naively, AI risks becoming a new kind of opaque authority, or simply a "magic wand" that produces persuasive but ungrounded judgments.

PoliScore is built to occupy the space between these extremes. It is not merely a scoring engine or a dashboard, but an attempt to:

- define a rigorous, academically grounded theory of policy quality,
- formalize a repeatable process for evaluating bills along that theory,
- translate those evaluations into sectoral and overall impact scores,
- aggregate impacts into transparent, interpretable legislator grades,
- and provide benchmark tasks for testing the AI systems used along the way.

The long-term vision is simple:

> *If we can show that we can evaluate bills in a non-partisan, well-grounded, and reproducible way, then we can evaluate legislators by aggregating the impacts of their bill-level interactions.*

PoliScore therefore has two tightly coupled layers:

1. a **policy quality layer**, which evaluates the structure, evidence base, and predicted societal impact of individual bills; and
2. a **legislator performance layer**, which aggregates those bill-level impacts according to how legislators sponsor, co-sponsor, and vote on them.

This white paper describes the theoretical foundations of PoliScore, its bill-level framework, the sectoral and overall impact model, the legislator aggregation logic, and the PoliBench benchmark suite used to validate AI systems in this domain. While the current implementation uses AI models as evaluators, the process is designed to be model-agnostic and could, in principle, be executed by human analysts following the same rubric.

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 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.

The rest of this paper develops both the theoretical and practical aspects of this discipline and shows how it can be extended from bill-level analysis to legislator-level performance grading.

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) 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.

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.

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 aggregate 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.

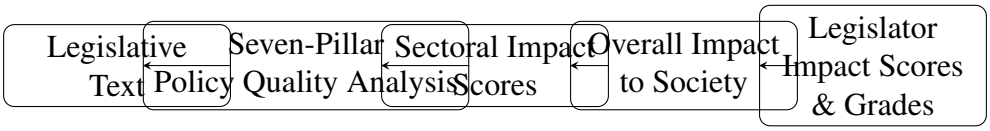

Figure 1: High-level PoliScore pipeline from bill text to legislator grades.

While this pipeline is currently implemented 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.

4.2 Bill-Level Scoring Prompt and Outputs

In its current 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.

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 raw impact score S_ℓ is defined as:

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

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.

Alternative weighting schemes can be explored (for example, increasing the weight of sponsorship relative to floor votes), but any change must be documented and kept constant for the period being analyzed.

4.5 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 S_ℓ. 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 S_ℓ.

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.

4.6 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.

4.7 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,
- 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

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 receives:

- 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 reveals the policy reasoning profile of an AI model and determines 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.

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 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.

6.3 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.

6.4 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 raw score S_ℓ as a normalized weighted sum of I_b^*.
4. Map S_ℓ to a letter grade via the standard threshold function.
5. Generate sectoral breakdowns of the legislator's performance (e.g., average impact in healthcare vs. energy).
6. 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.

6.5 Interpretability and Justification

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

6.5.1 Bill-Level Explanations

For each bill, PoliScore produces:

- 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.

6.5.2 Legislator-Level Explanations

For each legislator, PoliScore provides:

- the overall 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.

6.5.3 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.

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 framework complements 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. PoliScore:

- evaluates seven dimensions of policy quality rather than a single economic dimension;
- focuses on pre-implementation design soundness and sectoral societal impacts;

- assesses governance structure, feasibility, fairness, and systemic risk;
- aggregates bill impacts into legislator performance scores using transparent rules.

CBO answers: *"What will this cost?"*
PoliScore answers: *"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 answers: *"How structurally sound is this policy, and what is the aggregate 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 answers: *"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 complements 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 Contribution
CBO	Budgetary/fiscal impacts	Governance, feasibility, fairness, systemic risk, individual records	Provides holistic pre-implementation structural and impact evaluation, plus legislator aggregation.
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 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.

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 validates 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.

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.

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.

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.

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.

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 performance 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.
- **Civic integration.** Tools built on PoliScore could support voters, journalists, advocacy groups, and even legislators themselves in understanding and improving legislative performance.

10.2 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.

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.