AI Agents For Decision-Making in Climate Governance Using Policy Benchmarks

Shan Shan
Zhejiang University
shanshanfy@gmail.com

Abstract

Climate change governance requires navigating complex policy documents, including treaties, regulations, and socio-political frameworks. Understanding these texts is essential for evidence-based decision-making but remains challenging due to their complexity and domain specificity. This study explores the potential of AI agents to support policy reasoning and decision-making through structured evaluation on climate policy benchmarks, with a focus on dynamic governance scenarios. Drawing on global frameworks such as the UN Sustainable Development Goals (UNSDGs) and IPCC assessment pathways, this study evaluates agents using datasets such as Climate-FEVER (factual claim verification), LegalBench (legal reasoning), and PolicyQA (policy question answering). Target tasks include treaty interpretation, socio-political analysis, adaptation policy reasoning, and scenario-based planning. This study introduces a hybrid evaluation framework combining expert assessment and interdisciplinary feedback to systematically benchmark AI agents performance in climate governance, identifying their strengths, limitations, and potential for real-world support. It aims to bridge AI and climate governance, a tale of two systems, into a tale of collaboration.

1 Motivation

Benchmarks orient AI Liang et al. [2022]. Climate change governance demands coordinated and evidence-based action grounded in complex legal, regulatory, and socio-political contexts. Understanding and implementing effective policy requires navigating treaties, frameworks, and trade-offs across sectors and nations. With recent advances in AI, particularly the emergence of agentic systems built on large language models (LLMs), new opportunities arise for augmenting policy reasoning and decision-making. Traditionally the domain of experts in environmental law, political science, and economics, policy analysis can now be supported by AI agents capable of interactive reasoning, scenario simulation, and context-aware recommendationsDurante et al. [2024], Safaei and Longo [2024]. Evaluated through structured policy benchmarks, such agents offer a path toward more transparent, data-driven, and adaptive approaches to climate governance.

Is benchmarking the only effective bridge between AI and climate governance? Should we scale within large, evolving benchmark suites, or develop targeted, domain-specific modules tailored to climate policy tasks?

This study focuses on agentic performance, the behavior of systems built on LLMs powered AI agents within climate policy domains. GPT-5 is evaluated alongside GPT-4, BERT, and LegalBERT on tasks central to climate governance, including treaty interpretation, socio-political reasoning, and adaptation policy analysis. By framing climate policy as a reasoning benchmark, the analysis investigates the extent to which typical LLMs based agents can bridge the gap between technical expertise and policy-informed decision-making.

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2025.

A critical dimension of this evaluation involves automated fact-checking, given the importance of distinguishing accurate from misleading climate information in effective governance Lewandowsky [2021]. Using benchmark datasets such as Climate-FEVERDiggelmann et al. [2020], the study assesses the capacity of LLM-powered agents to validate climate-related claims, generating outputs that are transparent, context-aware, and timely. Overall, the findings aim to highlight both the potential and limitations of LLM-based agents in supporting evidence-based policymaking, climate communication, and governance workflows (see Figure 1).

This study contributes three key components:

- **Preliminary Hybrid Agent-based System.** The systematic framework includes agent collaboration across functional roles, automated reasoning, and a structured output pipeline. It aims to use general-purpose benchmarks for scale, and specilized shared tasks for policy depth. It is evaluated across specialized benchmarks, including Climate-FEVER (fact verification in climate discourse)Diggelmann et al. [2020], LegalBenchGuha et al. [2023] (legal and regulatory reasoning), and PolicyQA (question answering over policy documents).
- **Decision-making scenarios.** The analysis targets LLMs' domain-specific extensionsăperformance on treaty interpretation, socio-climate reasoning, and adaptation policy strategies to assess its utility in practical governance contexts.
- **Evaluation protocol.** A structured evaluation framework is proposed for assessing AI agents in climate policy reasoning, encompassing standardized task formats, benchmark datasets, and performance metrics spanning factual accuracy, legal soundness, and policy relevance.

2 What has been done

LegalBench has catalyzed an entire research direction in legal AI. From collaborative construction to large-scale benchmarking Guha et al. [2023], critical reflections Kapoor et al. [2024], and extensions into RAG and retrieval Pipitone and Alami [2024], Zheng et al. [2025], Hui et al. [2025], Sinha and Sharma [2025], Gupta et al. [2025], the work demonstrates both the promise and limitations of LLMs in legal contexts. It provides a critical foundation for future benchmarks in climate policy and environmental law, where similar challenges of technical language, statutory complexity, and high-stakes reasoning arise.

3 Methodology

Benchmarks. *ClimatePolicyQA* is the benchmark with 10,000 questions derived from IPCC reports [1], [2], [3],UNFCCC agreements and documents[4], and national adaptation plans. This database is set up for testing the comprehension of legal language, policy targets, and enforcement mechanisms. They are sourced from IPCC, UNFCC agreements. The resources provide specialized questions/answers to test models. *EnvLawBench, propsed* based on LawBenchFei et al. [2023] is a legal reasoning benchmark. It could derive from environmental statutes, regulations, and international treaties. It focus on evaluating statutory interpretation, conflict resolution, and legal reasoning. It is like LegalBechGuha et al. [2023], but more climate-speicific. *Mitigation & Adaptation Trade-off Corpus (MATC), proposed.* This is a scenario-based benchmark corpus. It curates a dataset of policy scenarios requiring reasoning about trade-offs, such as carbon pricing versus equity impacts, with a focus on reasoning about trade-offs and socio-economic impacts. This benchmark functions more as a reasoning and evaluation dataset than as a strict question-answering (QA) benchmark. *Sustainability Negotiation Simulation (SNS), proposed* is the benchmark in dialogue/ simulation dataset. It functions as role-playing dataset simulating multi-stakeholder negotiations, with multi-stakeholder negotiation scenarios. It test the models ability to generate persuasive, context-aware policy arguments, with focus on persuasion and argument generation in climate policy.

[1] `https://www.ipcc.ch/report/ar6/wg1/`

[2] `https://www.ipcc.ch/report/ar6/wg2/`

[3] `https://www.ipcc.ch/report/ar6/wg3/`

[4] `https://unfccc.int/documents`

Model-task alignment. This study employs multiple language models, each assigned to tasks aligned with their respective strengths. *GPT-5* is evaluated on active reasoning tasks, including simulating negotiations, generating policy recommendations, and interpreting complex climate governance frameworks. *GPT-4* is applied to text summarization, supporting the identification of broader policy trends. *BERT* is utilized for fact-checking and information extraction in climate discourse, such as verifying claims using the Climate-FEVER dataset. *LegalBERT* is used for legal interpretation tasks, including analysis of environmental statutes, regulatory texts, and treaties, as represented in benchmarks such as EnvLawBench.

Evaluation. The evaluation framework will employ a Mediator–Advocate mechanism for automated climate claim assessment. Advocates will represent perspectives derived from specific authoritative text sources, including the IPCC Intergovernmental Panel on Climate Change [2023], WMO World Meteorological Organization [2024], AbsCC (will contruct a corpus of abstracts from scientific climate literature), and 1000S (will contruct a corpus of abstracts authored by the top 1000 climate scientists) . In addition, GPT-5 will serve as a general-purpose advocate. For robustness analysis, will also include an adversarial advocate based on the NIPCC Idso et al. [2013] to represent contrarian or climate-denial perspectives.

4 Desiderate hybrid AI-agent-based system

Within this framework, AI agents act like Pac-Man, automatically gobbling up the messy, fragmented parts of the benchmark lifecycle, such as data updates, evaluations, and task coordination. This helps keep the system clean, adaptive, and continuously evolving across domains.

GPT-5, along with models such as GPT-4, BERT, and LegalBERT, is treated as an active reasoning agent capable of navigating climate governance with depth and coherence. Comparative evaluation across these models is designed to identify task-specific strengths, offering insights into how different AI systems can support effective, transparent, and evidence-based decision-making in climate policy. This work positions climate policy as a domain for benchmarking AI reasoning. It aims to contribute methodological innovations, including causal reasoning frameworks, and practical insights for climate change policy govenance. LLM-based agents' integration into this context presents the potential to act as a policy co-pilot, supporting globally coordinated and scientifically grounded climate governance.

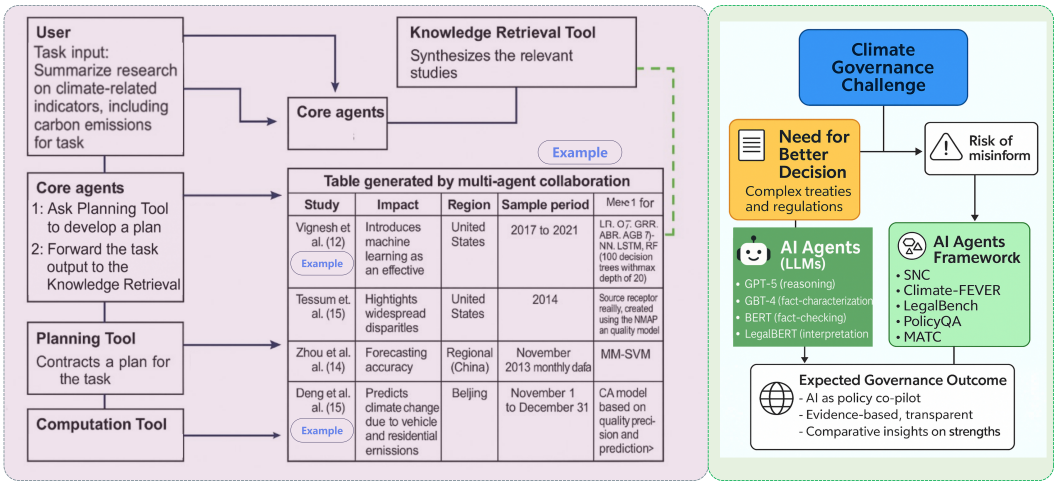

Figure 1: AI Agents and Benchmarking Frameworks for Evidence-Based Climate Governance

5 Discussion and limitations

This study investigates the role of AI agents in climate governance by evaluating their reasoning capabilities across benchmark tasks grounded in legal, policy, and scientific texts. Rather than focusing on a single model, the analysis highlights different language-model-based agents–such as

GPT-5, GPT-4, BERT, and LegalBERT–offer complementary strengths for tasks including treaty interpretation, fact verification, socio-political reasoning, and legal analysis.

Beyond task-specific capabilities, this work explores how AI agents can automate, orchestrate, and mediate key pain points across the benchmark lifecycle. Agents can translate between benchmark formats, wrap APIs, and synchronize tasks across large-scale evaluations (e.g., HELM Liang et al. [2022], BIG-BenchSuzgun et al. [2022]), helping reduce integration complexity. To manage evaluation costs, agents can selectively test or simulate policy scenarios. For governance fragmentation, they can maintain compatibility layers across frameworks such as the UNFCCC, IPCC, and national policies.

Together, these agentic capabilities point toward a scalable, adaptive infrastructure for climate policy benchmarking. It supports transparent, interdisciplinary, and evidence-based decision-making in global climate governance.

Nevertheless, despite being the latest model, LLMs are still in a testing phase and continues to exhibit flaws WIRED editors [2025], Reddit users [2025], Masood [2025]. It may misinterpret ambiguous legal language, overgeneralize from precedents, and lack grounding in non-textual climate data such as emissions statistics. Technical challenges are also evident, including unstable API performance under complex multi-turn interactions, incomplete reproducibility of outputs, and inconsistent logical reasoning across scenarios. Ethical concerns further complicate deployment, particularly regarding model bias, potential misuse of information, and the urgent need for transparent oversight.

6 Conclusion

This study examines how AI research can be bridged with global climate governance through the use of policy benchmarks. It highlights the value of using general-purpose benchmarks for scalability and specialized shared tasks for policy-specific depth. The study proposes a hybrid architecture in which AI agentsbuilt on large language modelssupport climate governance through structured reasoning over benchmarked tasks. This framework leverages the complementary strengths of GPT-5, GPT-4, BERT, and LegalBERT across tasks such as treaty interpretation, fact verification, legal analysis, and socio-political reasoning.

Beyond performance evaluation, the study explores how agents can streamline the benchmark lifecycle. Agents can translate formats, wrap APIs, and synchronize tasks across benchmark suites; selectively simulate scenarios to reduce evaluation costs; and maintain compatibility across governance frameworks like the UNFCCC and IPCC. They can also curate tasks, update data, and coordinate contributor workflowsenabling the development of living, scalable benchmarks for climate policy.

These contributions position AI agents in a Pac-Man-like role, cleaning up the messy parts of the system, as infrastructure for adaptive, evidence-based, and globally coordinated decision-making in climate governance; bridging AI and climate governance, a tale of two systems, into a tale of collaboration.

7 Ackowledgement

The author thanks NeurIPS reviewers for the feedback on model selection, and Prof. Manmeet Singh and Dr. Austin Naveen Sudharsan for early comments.

References

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*, 2020.

Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, et al. Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568*, 2024.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*, 2023.

Nikhil Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Zhiwei Li, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 44123–44279, 2023.

Jatin Gupta, Ananya Sharma, Sarthak Singhania, and Ali Imran Abidi. Legal assist ai: Leveraging transformer-based model for effective legal assistance. *arXiv preprint arXiv:2505.22003*, 2025.

Zhen Hui, Yan Rong Dong, Ehsan Shareghi, and Nigel Collier. Trident: Benchmarking llm safety in finance, medicine, and law. *arXiv preprint arXiv:2507.21134*, 2025.

C. D. Idso, R. M. Carter, and S. F. Singer, editors. *Climate Change Reconsidered II: Physical Science*. Heartland Institute, Chicago, IL, 2013.

Intergovernmental Panel on Climate Change. *Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. IPCC, Geneva, Switzerland, 2023.

Sayash Kapoor, Peter Henderson, and Arvind Narayanan. Promises and pitfalls of artificial intelligence for legal applications. *arXiv preprint arXiv:2402.01656*, 2024.

Stephan Lewandowsky. Climate change disinformation and how to combat it. *Annual review of public health*, 42(1):1–21, 2021.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Adnan Masood. Openai's gpt-5 is here: A deep dive into the ai that's , 2025. URL `https://medium.com/@adnanmasood/openais-gpt-5-is-here`. Accessed: Aug 2025.

Niccolò Pipitone and Ghadah H. Alami. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint arXiv:2408.10343*, 2024.

Reddit users. Discussion on gpt-5 performance compared to gpt-4.1. `https://www.reddit.com/r/OpenAI`, 2025. Accessed: Aug 2025, community thread with 120+ comments.

Mehrdad Safaei and Justin Longo. The end of the policy analyst? testing the capability of artificial intelligence to generate plausible, persuasive, and useful policy analysis. *Digital Government: Research and Practice*, 5(1):1–35, 2024.

Divya Sinha and Omkar Sharma. Generating legal arguments using llm and vector database to support precedents. In *2025 International Conference on Next Generation Information System Engineering (NGISE)*, volume 1, pages 1–5. IEEE, March 2025.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

WIRED editors. Developers say gpt-5 is a mixed bag, 2025. URL `https://www.wired.com/story/openais-gpt-5-is-here`. Accessed: Aug 2025.

World Meteorological Organization. *State of the Global Climate 2024*. WMO, Geneva, Switzerland, 2024.

Long Zheng, Nikhil Guha, Jamshid Arifov, Shiyu Zhang, Marios Skreta, Christopher D. Manning, and Daniel E. Ho. A reasoning-focused legal retrieval benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law*, pages 169–193, March 2025.