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

Shan Shan

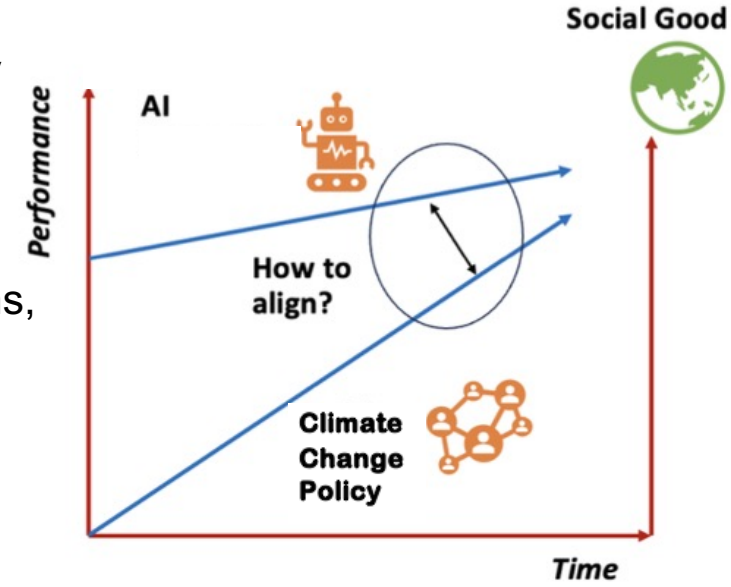
Zhejiang University
shanshanfy@gmail.com

Outline

- A Tale of Two Systems: AI and Climate Governance
- Benchmarking is the Bridge?
- Automating Climate Policy Analysis with Socio-economic-aware Multimodal Multiagent AI
- A Tale of Collaboration

A Tale of Two Systems

- **AI Matters for Climate Governance**: models policy outcomes, negotiates trade-offs, scales decision support.
- **Climate Governance Matters for AI**: sets the norms, values, and deployment constraints for AI use.
- **Societal impact** depends on their **alignment.**



Benchmarks orient AI

- “Benchmarks orient AI” (Liang *et al.*, 2023)
- Bridging AI research and global climate governance through policy benchmarks
- Growth on the existing larger living benchmark (like HELM)? [Path 1]
- Or independent domain-specific modules growth to be targeted, and cost-efficient way? [Path 2]

Path 1: Extensibility

- Bridging AI research and global climate governance through policy benchmarks

  - Modular structure: new tasks, languages, or document types can be added
  - Foundation for a public **leaderboard** and open-source **evaluation toolkit**

Path 2: Specificity

- Use general-purpose benchmarks for scale, and specialized shared tasks for policy depth?

Potential scenarios	Metrics(Current Dataset Example)	Purpose
Factual Verification	**Climate-FEVER**	Verifies claims against evidence
Legal Entailment	**LegalBench**	Determines obligation/permission
Policy QA	**PolicyQA**	Answers structured queries over long docs
Simulation & Negotiation	**SNC, MATC**	Multiagent treaty reasoning tasks

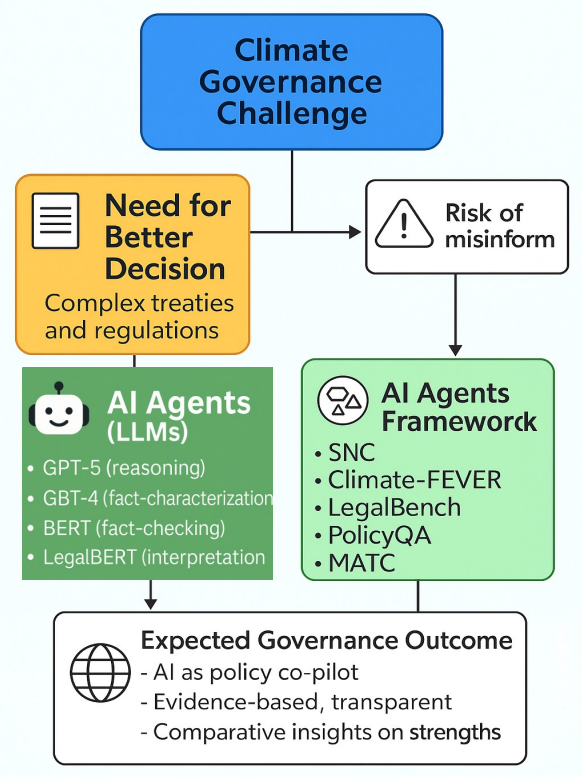
Best of Both Worlds?

- **AI agents** can **automate**, **orchestrate**, and **mediate** many pain points in benchmark lifecycle tasks that are otherwise manual, fragmented, or expensive.
- AI agents act like **Pac-Man gobbling up** complexities.

Challenge	Solution with AI Agents
Complexity of integrating with large existing benchmarks	Use agents to **translate formats**, **wrap APIs**, or **synchronize tasks** between selected benchmark and existing ones (e.g., HELM, BIG-Bench).
High cost of full-scale evaluation	Deploy agents to **selectively evaluate**, run **active testing**, or **simulate policy scenarios**, reducing overhead.
Fragmentation across institutions and governance models	Agents can maintain compatibility layers for different governance frameworks (UNFCCC, IPCC, local policies), publishing results in shared policy evaluation formats to ensure cross-institutional interoperability.
Maintaining a living benchmark	Use autonomous agents to **curate new tasks**, **update data**, or **run evaluations** on new models.
Coordinating multiple contributors	AI agents can act as **orchestrators**, managing tasks like review queues, annotation quality, or evaluation status.

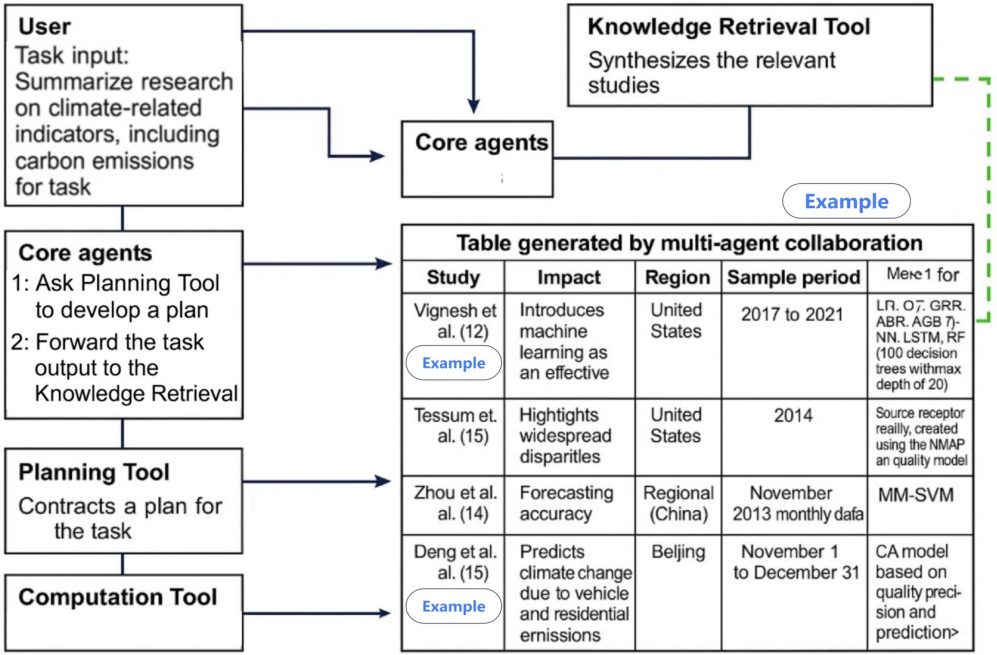
Preliminary Multi-modal Agents & Models

- **Core modules** are shared and compatible with existing benchmarks.
- **Domain-specific extensions** (like climate policy tasks) are added as plug-ins.
- **Agents** manage updates, evaluations, and curation semi-autonomously.



Desiderate Hybrid Agent-based System

- Agent collaboration across functional roles; Automated reasoning; Structured output pipeline.



Challenges: Interpreting, Summarizing, and Reasoning

- Different agent-functions need to be **coordinated** well across tasks (*Reflexive Multimodal Learning*).
- **Reasoning** (causal) is still challenging.
- LLMs struggle with legal, numeric, and policy complexity.

A tale of collaboration

1. Reflexive AI for the Climate Change Roadmap: compile development and implementation, empirical validation and societal implications, integration with existing AI technology.

2. Project location : https://github.com/shanshanfy/TowardsReflexiveAI

Thank You