

# Al Agents For Decision-Making in Climate Governance Using Policy Benchmarks

**Shan Shan** 

Zhejiang University, shanshanfy@gmail.com



# From: A Tale of Two Systems

#### Climate Change → Al

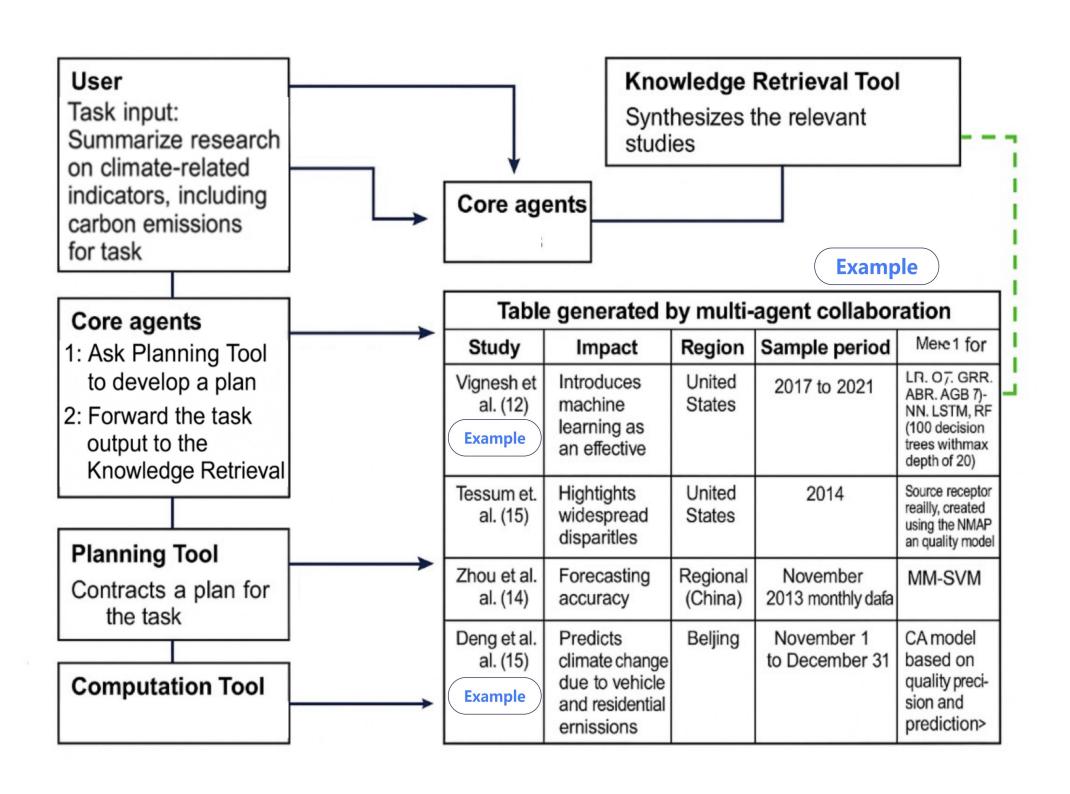
- Al and climate change are a tale of two systems, how to bridge them?
- Climate Change: A Global Governance Challenge; urgent, informed, and coordinated action is required.
- Policymaking involves complex legal frameworks, global treaties, and socio-economic trade-offs.

#### **Al**→ Climate Change

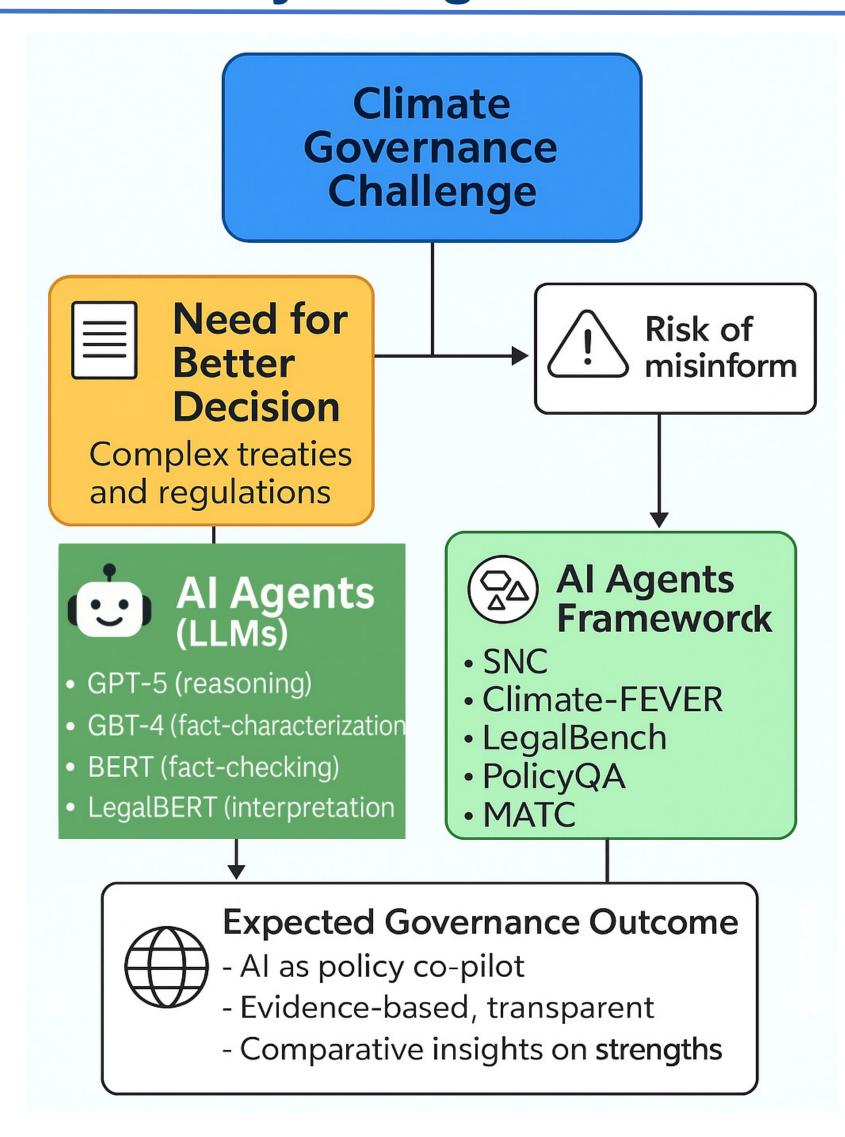
- Benchmarks orient Al
- Is benchmarking the only effective bridge between Al and climate governance?
- Should we scale within large, evolving benchmark suites, or develop targeted, domain-specific modules tailored to climate policy tasks?

## Methodology

#### Benchmarks, model-task alignment, Evaluation



### **Hybrid Policy-Al Agent Framework**



# **Policy Benchmark**



Use generalpurpose benchmarks for scale, and specialized shared tasks for policy depth?

### 1. Benchmarking LLMs

- Climate-FEVER (fact verification)
- LegalBench (legal reasoning)
- PolicyQA (policy document QA)

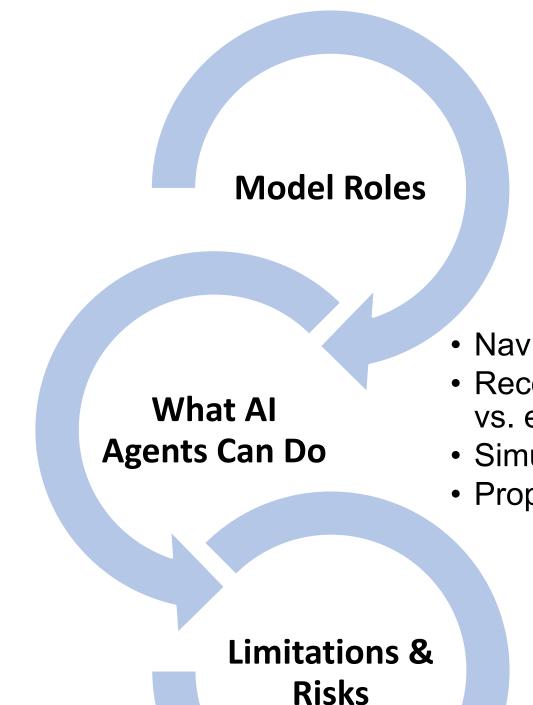
### 2. Decision-Making Scenarios

- Treaty reasoning
- Climate adaptation planning
- Socio-political trade-off analysis

### 3. Evaluation Protocol

Accuracy, legal soundness, policy relevance

## **Capabilities and Limitations**



- GPT-5: Leads in simulating negotiations & policy reasoning
- GPT-4, BERT, LegalBERT: Support tasks like factchecking & legal text interpretation

Navigate dense legal/policy documents

- Reconcile competing goals (e.g., economy vs. emissions)
- Simulate stakeholder perspectives
- Propose new governance strategies

Misreads ambiguous legal languageOvergeneralizes precedents

- Struggles with non-textual data (e.g., climate metrics)
- Technical issues: unstable APIs, output inconsistency
- Ethical concerns: bias, lobbying misuse, oversight gaps

## **To:** A Tale of Collaboration

#### **Scale and Benchmark**

This work suggests that integrating AI active agents, including GPT-5, GPT-4, BERT, and LegalBERT, into climate governance can provide valuable insights across different aspects of decision-making, from policy interpretation to fact-checking and legal analysis.

#### **Future Research**

By understanding the complementary strengths of these models, we can accelerate evidence-based, equitable, and globally coordinated action against climate change.

#### Acknowledgement

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











