

---

# Reflexive Evidence-Based Multimodal Learning for Clean Energy Transitions: Causal Insights on Cooking Fuel Access, Urbanization, and Carbon Emissions

---

Shan Shan  
Zhejiang University  
shanshanfy@gmail.com

## Abstract

Achieving Sustainable Development Goal 7 (Affordable and Clean Energy) requires not only technological innovation but also a deeper understanding of the socio-economic factors that influence energy access and carbon emissions. Despite growing attention to these drivers, key questions remain, particularly regarding how to quantify socio-economic impacts, how these impacts interact across domains such as policy, technology, and infrastructure, and how feedback processes shape energy systems. To address these gaps, this study introduces ClimateAgents, an AI-based framework that combines large language models with domain-specialized agents to support hypothesis generation and scenario exploration. Leveraging 20 years of socio-economic and emissions data from 265 economies, countries and regions, and 98 indicators drawn from the World Bank database, the framework applies a machine learning based causal inference approach to identify key determinants of carbon emissions in an evidence-based, data-driven manner. The analysis highlights three primary drivers: (1) access to clean cooking fuels in rural areas, (2) access to clean cooking fuels in urban areas, and (3) the percentage of population living in urban areas. These findings underscore the critical role of clean cooking technologies and urbanization patterns in shaping emission outcomes. In line with growing calls for evidence-based AI policy, ClimateAgents offers a modular and reflexive learning system that supports the generation of credible and actionable insights for policy. By integrating heterogeneous data modalities, including structured indicators, policy documents, and semantic reasoning, the framework contributes to adaptive policymaking infrastructures that can evolve with complex socio-technical challenges. This approach aims to support a shift from siloed modeling to reflexive, modular systems designed for dynamic, context-aware climate action.

Strengthening science-informed, evidence-based approaches to AI policy is increasingly essential[1]. In domains such as clean energy and climate change, the question is how to integrate the strengths of narrow AI components into a unified, context-aware agent, one able to process multimodal data, reason causally, interact socially, and adapt continuously to evolving policy challenges. Recent advances in artificial intelligence (AI), particularly large language models (LLMs), present new opportunities for sustainability research. LLMs are being applied in environmental science, economics, and climate modeling [2, 3, 4, 5, 6, 7, 8], with transformer-based architectures enabling reasoning, inference, and policy analysis [9, 10]. They also support hypothesis generation, in-context retrieval, and multimodal data synthesis [11, 12]. However, sustainability applications remain limited, often constrained to text summarization or keyword-based mapping [13, 14, 15, 16, 17]. Current LLMs lack integration with structured data, causal inference, and adaptation to new socio-environmental indicators [18]. A key barrier is the absence of causal reasoning datasets: while general-purpose corpora from Google [19], Bing [20], and user interactions (e.g., ShareGPT, WildChat) [21] exist, few

support causal inquiry in LLM prompts [22, 23]. Critically, no dataset addresses socio-climate-related causal questions, leaving a major gap for advanced applications [24].

To ensure reliable and standardized evidence-based analysis, this research adopts the World Bank Development Indicators—a widely recognized, high-quality, and publicly available data framework. This data-driven foundation enhances the credibility and precision of the study’s outputs. By applying causal inference techniques with machine learning algorithms, the analysis moves beyond simple correlation to uncover deeper, more robust relationships. This enables more grounded, interpretable reasoning for policy-making. Furthermore, the integration of large language models (LLMs) supports evidence-based analysis by generating outputs that aim to be *credible and actionable*, as their interpretability can facilitate context-aware and informed decision-making [25, 26, 27].

The orchestration of these three modules reflects a system-level design philosophy rooted in modularity, specialization, and agent-based coordination. To address these challenges and extend the utility of large language models (LLMs), this work proposes a multi-agent architecture grounded in Minsky’s philosophy of modular, emergent intelligence. Rather than treating LLMs as monolithic tools, the proposed system distributes reasoning and task execution across a set of interacting agents, each specialized for distinct functions. The resulting framework—ClimateAgents—is a reflexive, causal modeling system powered by GPT family models [28] and accessed via the OpenAI Application Programming Interface [29]. It moves beyond static prompt-response paradigms, enabling adaptive reasoning within complex socio-environmental systems. Central to this architecture is the concept of *Reflexive* machine learning, defined here as a process through which agents iteratively adjust their prompts, inference strategies, or actions in response to environmental feedback and task complexity, thereby supporting context-aware and adaptive decision-making.

The contributions of this work are as follows: (i) introduction of a reflexive multiagent architecture for causal analysis and policy simulation in socio-environmental contexts; (ii) integration of multimodal data with LLM-driven agents to complement statistical models through simulation, reasoning, and hypothesis generation; and (iii) proposal of Reflexive Machine Learning as a natural-language interface for interpretable modeling of complex systems.

## Methods

### ClimateAgents

ClimateAgents consists of three components: (i) a perception layer that structures multimodal inputs into formal representations (e.g., indicators, semantic frames); (ii) a reasoning layer for planning, inference, and adaptive decision-making; and (iii) an operation layer that performs causal inference, modeling, and policy simulations, with outputs interpreted via LLMs. A continuous agent feedback loop enables real-time refinement and contextual adaptation for evidence-based policy support (Figure 1, 3).

### Causal Inference

Based on former work[8], this study furhter introduces a *three-stage comparative framework* for investigating causal relationships in the context of social science and climate change, aimed at supporting evidence-based reasoning. The pipeline combines (i) correlation analysis to identify initial statistical associations, (ii) machine learning based causal discovery to estimate structural dependencies, and (iii) LLM-guided prompt exploration to surface contextual explanations and generate policy-relevant hypotheses. Each stage contributes distinct but complementary evidence toward causal interpretation, facilitating more transparent and informed downstream analysis. This approach is designed to support the development of empirical insights that can inform decision-making in complex socio-environmental systems.

### Evidence Informed Policy Reasoning

**Evidence retrieval** is demonstrated through text classification of agent-generated prompts (Figure 2), which revealed themes related to carbon emission prediction, including model diversity, geographic specificity, and environmental justice. Using Biopython and the NCBI Entrez database, the system efficiently retrieves and synthesizes relevant literature, supporting large-scale climate and air quality

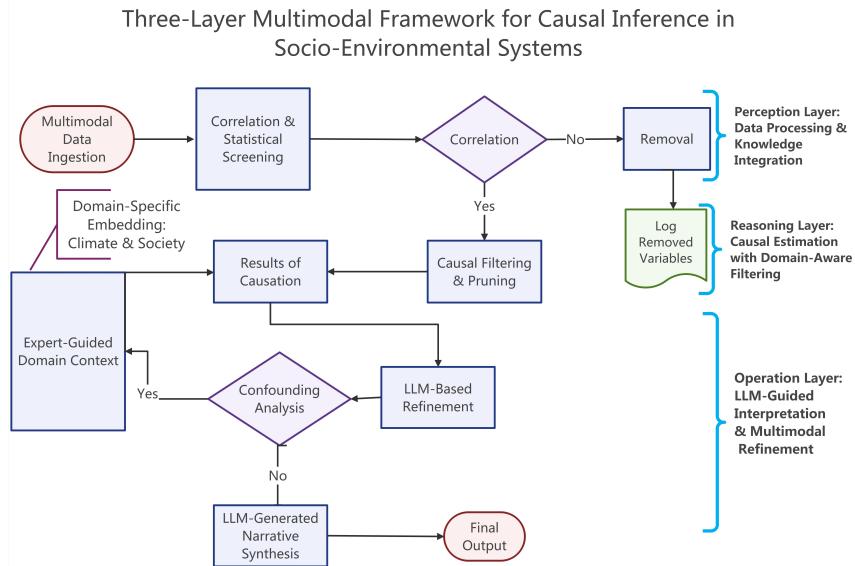


Figure 1: Reflexive Multimodal ClimateAgents Framework

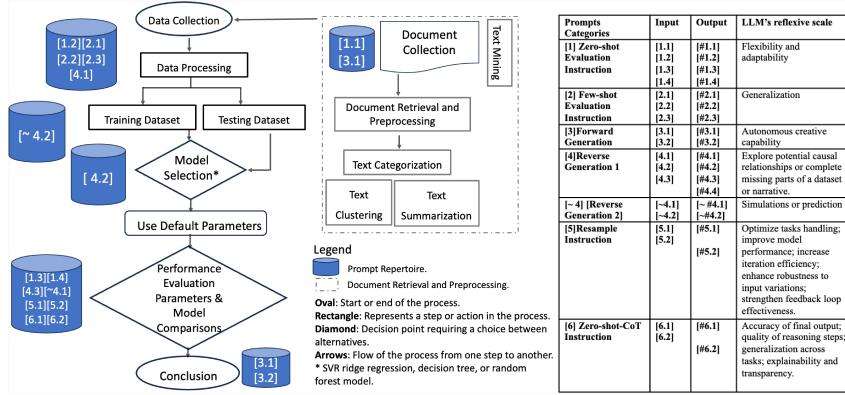


Figure 2: Perception Layer Framework for Climate Change Data Processing and Model Evaluation.

research. **Causal effects** are estimated following Rolland et al. [30], modeling each variable as a function of its causal parents with additive noise. Leaf nodes are identified using score function derivatives, and topological ordering is achieved by sequential leaf removal, with the Jacobian approximated by the Stein gradient estimator and refined through the CAM procedure [30]. **Validation and interpretation** involve domain expertise and standard metrics, with results highlighting key drivers such as rural and urban access to clean fuels and urbanization growth. To validate LLMs contribute to causal inference, the framework applies World Bank Development Indicators and employs a taxonomy of causality [31, 32] grounded in Pearl’s Causal Hierarchy [33, 34]. This taxonomy distinguishes direct, preventative, facilitative, resultative, and influential causal verbs, ensuring LLM inquiries classify relations accurately and avoid confounding [35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46]. Data on carbon emissions per capita from 265 economies, countries and regions, covering the period from 2000 to 2020, are sourced from Climate Watch Historical GHG Emissions.

## Results and Discussion

The multi-agent system initiated the analysis using a planning tool that decomposed complex user queries into actionable subtasks. Execution incorporated multiple tools, including automated statistical computation, literature analysis, knowledge retrieval from scientific databases, and code generation for downstream modeling (Figure 2, 3).

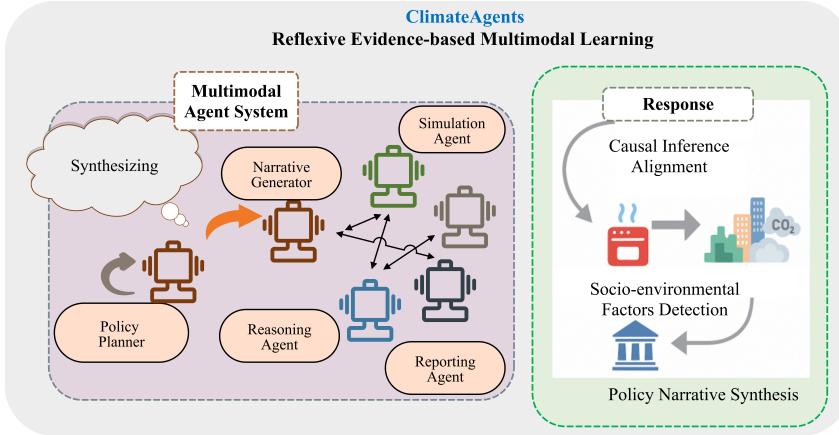


Figure 3: Compact multi-agent framework for policy scenario modeling. Each agent fulfills a specialized role.

This study identified three key socioeconomic variables with strong causal effects on per capita carbon emissions: (i) ***Access to clean fuels and technologies for cooking, rural (% of rural population)*** (EG.CFT.ACCTS.RU.ZS), which showed a strong negative relationship with emissions, as higher access reduces biomass reliance and climate impact in less urbanized regions [47, 48]; (ii) ***Access to clean fuels and technologies for cooking, urban (% of urban population)*** (EG.CFT.ACCTS.UR.ZS), which similarly lowered emissions in dense populations through improved energy efficiency [49]; and (iii) ***Urban population (% of total population)*** (SP.URB.TOTL.IN.ZS), which was positively associated with emissions, reflecting increased demand for transport, infrastructure, and industry [50, 51, 52]. These results were derived from a causal analysis of 20 years of data across 265 economies, countries and regions, covering 98 socioeconomic indicators. Full steps are provided in SI Appendix.

To evaluate this capacity, the system applied instruction-based prompting techniques, including zero-shot, few-shot, and chain-of-thought (CoT) prompts [53, 9, 54]. Prompts were anchored to previously identified causal variables—access to clean fuels in rural and urban contexts (EG.CFT.ACCTS.RU.ZS and EG.CFT.ACCTS.UR.ZS) and urban population percentage (SP.URB.TOTL.IN.ZS)—and structured using a five-part causal taxonomy from the CausalQuest framework [24], aligned with Pearl’s Causal Hierarchy [33, 34]. The five causal query types—Direct, Preventative, Facilitative, Resultative, and Influential—were designed to elicit different dimensions of causal reasoning [31, 32]. An illustrative output is shown in Figure 4, where the LLM agent (“AI Social-Scientist”) was prompted to assess whether disparities in clean fuel access might impact the effectiveness of emission-reduction policies. The model responded by highlighting the role of infrastructure inequality, policy generalization limits, and the importance of regional context in designing equitable interventions.

<b>User:</b> Rural areas (EG.CFT.ACCTS.RU.ZS) have lower access to clean fuels than urban ones (EG.CFT.ACCTS.UR.ZS), and urbanization (SP.URB.TOTL.IN.ZS) is increasing. Could variability in clean fuel access affect the effectiveness of environmental policies at reducing emissions?
<b>AI Social-Scientist (GPT-4-turbo): Policy Effectiveness:</b> Greater variability weakens uniform outcomes. <b>Emission Impact:</b> Urban-focused policies may yield quick results, but lasting success requires rural inclusion. <b>Urbanization Factor:</b> Shifts where and how policies work; less urbanized regions need tailored approaches. <b>Policy Design:</b> Reducing variability can support equitable and impactful strategies.

Figure 4: Conversation between a user and an AI social-scientist model (GPT-4-turbo) on clean fuel access and policy effectiveness.

## Summary and Future Perspectives

This study introduces **ClimateAgents**, a multi-agent framework that integrates large language models with a data-driven approach for climate policy analysis. Structured into perception, reasoning, and operation layers, the system combines planning, reasoning, simulation, narrative, and reporting

agents to model complex socio-environmental dynamics. By leveraging the reasoning and synthesis capabilities of LLMs alongside computational efficiency, ClimateAgents enables flexible analysis of clean energy, urbanization, and carbon emissions, integrating heterogeneous datasets into interpretable narratives and visualizations that enhance accuracy and support evidence-based policy.

Nonetheless, limitations remain. Causal inference depends on data quality, simplified assumptions about emissions may not generalize, and automated script generation for complex simulations is not yet feasible. Causal inquiry in LLMs is also shaped by social norms and linguistic context [55, 32]. Future integration of advanced foundation models and generative tools could enhance scalability, reduce computational costs, and expand predictive depth, establishing ClimateAgents as a flexible platform for intelligent climate policy simulation.

## 1 Acknowledgement

The author gratefully acknowledges Prof. Fei Wu for his guidance on the causal framework, Dr. Anpeng Wu for his support with causal analysis, and Prof. Hao Chen for his direction in developing the multimodal framework.

## References

- [1] Rishi Bommasani, Sanjeev Arora, Jennifer Chayes, Yejin Choi, Mariano-Florentino Cuéllar, Li Fei-Fei, Daniel E Ho, Dan Jurafsky, Sanmi Koyejo, Hima Lakkaraju, et al. Advancing science-and evidence-based ai policy. *Science*, 389(6759):459–461, 2025.
- [2] Muhammad Usman Hadi, Qasem Al Tashi, Abbas Shah, Rizwan Qureshi, Amgad Muneer, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, et al. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*, 2024.
- [3] Marcel Binz, Stephan Alaniz, Adina Roskies, Balazs Aczel, Carl T. Bergstrom, Colin Allen, Daniel Schad, and et al. How should the advancement of large language models affect the practice of science? *Proceedings of the National Academy of Sciences*, 122(5):e2401227121, 2025.
- [4] Matthias C. Rillig, Marlene Ågerstrand, Mohan Bi, Kenneth A. Gould, and Uli Sauerland. Risks and benefits of large language models for the environment. *Environmental science & technology*, 57(9):3464–3466, 2023.
- [5] Mohaimenul Azam Khan Raian, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE access*, 12:26839–26874, 2024.
- [6] Kasra Motlaghzadeh, Ali Eyni, Massoud Behboudian, Parastoo Pourmoghim, Saeed Ashrafi, Reza Kerachian, and Keith W. Hipel. A multi-agent decision-making framework for evaluating water and environmental resources management scenarios under climate change. *Science of The Total Environment*, 864:161060, 2023.
- [7] Jörg P Müller and Klaus Fischer. Application impact of multi-agent systems and technologies: A survey. *Agent-Oriented Software Engineering: Reflections on Architectures, Methodologies, Languages, and Frameworks*, pages 27–53, 2014.
- [8] Shan Shan. From correlation to causation: Understanding climate change through causal analysis and llm interpretations. *arXiv preprint arXiv:2412.16691*, 2024.
- [9] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022.
- [10] Zhijing Jin, Zichao Peng, Tushar Vaidhya, Bernhard Schoelkopf, and Rada Mihalcea. Mining the cause of political decision-making from social media: A case study of covid-19 policies across the us states. In *Findings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, 2021. Association for Computational Linguistics.

- [11] Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, et al. Many-shot in-context learning. *Advances in Neural Information Processing Systems*, 37:76930–76966, 2024.
- [12] Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. In-context learning with retrieved demonstrations for language models: A survey. *arXiv preprint arXiv:2401.11624*, 2024.
- [13] Yang Zhang, Hanlei Jin, Dan Meng, Jun Wang, and Jinghua Tan. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*, 2024.
- [14] Nazanin Firoozeh, Adeline Nazarenko, Fabrice Alizon, and Béatrice Daille. Keyword extraction: Issues and methods. *Natural Language Engineering*, 26(3):259–291, 2020.
- [15] Jinghan Yang, Hongjin Jin, Raphael Tang, Xu Han, Qi Feng, Haoyang Jiang, Shijie Zhong, Baobao Yin, and Xinyu Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 18(6):1–32, 2024.
- [16] Yotam Wolf, Nadav Wies, Omer Avnery, Yoav Levine, and Amnon Shashua. Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*, 2023.
- [17] Pablo Villalobos, Alex Ho, Jack Sevilla, Tim Besiroglu, Lukas Heim, and Maarten Hobbhahn. Will we run out of data? limits of llm scaling based on human-generated data. *arXiv preprint arXiv:2211.04325*, 2022.
- [18] Christopher A. Bail. Can generative ai improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121, 2024.
- [19] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Dave Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Michael Kelcey, Matthew-W. Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.
- [20] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human generated machine reading comprehension dataset. In Tarek R. Besold, Antoine Bordes, Artur S. d’Avila Garcez, and Greg Wayne, editors, *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches 2016 Co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016)*, volume 1773 of *CEUR Workshop Proceedings*, 2016.
- [21] Weizhi Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yue Deng. (inthe)wildchat: 570k chatgpt interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024.
- [22] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder Askell, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155, 2022.
- [23] Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation?, 2024.
- [24] Roberto Ceraolo, Dmitrii Kharlapenko, Amélie Reymond, Rada Mihalcea, Mrinmaya Sachan, Bernhard Schölkopf, and Zhijing Jin. Causalquest: Collecting natural causal questions for ai agents, 2024.
- [25] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

- [26] Sarah Wiegreffe and Ana Marasović. Teach me to explain: A review of datasets for explainable natural language processing. *arXiv preprint arXiv:2102.12060*, 2021.
- [27] Rishi Bommasani. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- [28] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [29] OpenAI. Openai api. <https://openai.com/blog/openai-api>. Accessed: 2025-3-10.
- [30] Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, pages 18741–18753. PMLR, 2022.
- [31] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2023.
- [32] Shaobo Cui, Zhijing Jin, Bernhard Schölkopf, and Boi Faltings. The odyssey of commonsense causality: From foundational benchmarks to cutting-edge reasoning, 2024.
- [33] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York, NY, first edition edition, May 2018.
- [34] Elias Bareinboim, Juan D. Correa, David Ibeling, and Thomas Icard. On pearl’s hierarchy and the foundations of causal inference. In Hector Geffner, Rina Dechter, and Joseph Y. Halpern, editors, *Probabilistic and Causal Inference: The Works of Judea Pearl*, volume 36 of *ACM Books*, pages 507–556. ACM, 2022.
- [35] Roxana Girju and Dan Moldovan. Mining answers for causation questions. In *Proc. The AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases*, pages 67–82, 2002.
- [36] Zornitsa Kozareva. Cause-effect relation learning. In *Workshop Proceedings of TextGraphs-7: Graph-based Methods for Natural Language Processing*, pages 39–43, 2012.
- [37] Mehwish Riaz and Roxana Girju. Recognizing causality in verb-noun pairs via noun and verb semantics. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 48–57, 2014.
- [38] Erika Nazaruka. An overview of ways of discovering cause-effect relations in text by using natural language processing. In *Evaluation of Novel Approaches to Software Engineering: 14th International Conference, ENASE 2019, Heraklion, Crete, Greece, May 4–5, 2019, Revised Selected Papers 14*, pages 22–38. Springer, 2020.
- [39] Christopher Allen. *A local grammar of cause and effect: A corpus-driven study*. PhD thesis, University of Birmingham, 2005.
- [40] Gill Harvey, Alison Loftus-Hills, Jo Rycroft-Malone, Angie Titchen, Alison Kitson, Brendan McCormack, and Kate Seers. Getting evidence into practice: the role and function of facilitation. *Journal of advanced nursing*, 37(6):577–588, 2002.
- [41] Phillip Wolff. Direct causation in the linguistic coding and individuation of causal events. *Cognition*, 88(1):1–48, 2003.

- [42] Hans Christian Boas. *Resultative constructions in English and German*. The University of North Carolina at Chapel Hill, 2000.
- [43] Maria Sandra Pena Cervel. A constructionist approach to causative frighten verbs. *Linguistics*, 53(6):1247–1302, 2015.
- [44] Albert S Yee. The causal effects of ideas on policies. *International organization*, 50(1):69–108, 1996.
- [45] Paul Slovic, Melissa L Finucane, Ellen Peters, and Donald G MacGregor. The affect heuristic. *European journal of operational research*, 177(3):1333–1352, 2007.
- [46] Paul Slovic, Melissa L Finucane, Ellen Peters, and Donald G MacGregor. Risk as analysis and risk as feelings: Some thoughts about affect, reason, risk and rationality. In *The feeling of risk*, pages 21–36. Routledge, 2013.
- [47] Solomon Prince Nathaniel and Ngozi Adeleye. Environmental preservation amidst carbon emissions, energy consumption, and urbanization in selected african countries: implication for sustainability. *Journal of Cleaner Production*, 285:125409, 2021.
- [48] Pramit Verma, Tanu Kumari, and Akhilesh Singh Raghubanshi. Energy emissions, consumption and impact of urban households: A review. *Renewable and Sustainable Energy Reviews*, 147:111210, 2021.
- [49] Muhammad Abubakr Naeem, Michael Appiah, John Taden, Richard Amoasi, and Bright Akwasi Gyamfi. Transitioning to clean energy: Assessing the impact of renewable energy, bio-capacity and access to clean fuel on carbon emissions in oecd economies. *Energy Economics*, 127:107091, 2023.
- [50] Steve Hankey and Julian D Marshall. Impacts of urban form on future us passenger-vehicle greenhouse gas emissions. *Energy Policy*, 38(9):4880–4887, 2010.
- [51] Reinhard Madlener and Yasin Sunak. Impacts of urbanization on urban structures and energy demand: What can we learn for urban energy planning and urbanization management? *Sustainable Cities and Society*, 1(1):45–53, 2011.
- [52] Ke Li and Boqiang Lin. Impacts of urbanization and industrialization on energy consumption/co2 emissions: does the level of development matter? *Renewable and Sustainable Energy Reviews*, 52:1107–1122, 2015.
- [53] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.
- [54] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.
- [55] Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online, November 2020. Association for Computational Linguistics.
- [56] Sayash Kapoor, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, et al. On the societal impact of open foundation models. *arXiv preprint arXiv:2403.07918*, 2024.

## Supplementary Information

### Knowledge Retrieval

Text classification of agent-generated prompts ([1.1], [3.1] of Figure 2 revealed key themes related to carbon emission prediction. The extracted summaries highlighted model diversity, geographic specificity, environmental justice considerations, and challenges in temporal and spatial scaling. The study utilizes the Biopython library and the NCBI Entrez database to provide an example of literature retrieval and analysis. Running this code snippet allows for efficient retrieval of article details, including titles, from the PubMed database, facilitating in-depth analysis and synthesis of research findings. It offers a tool for researchers investigating the complex relationship between climate change and global air quality, enabling the retrieval and analysis of a large volume of articles. These results demonstrate the system's ability to identify cross-cutting concerns in climate-related modeling efforts.

### Causal Effects Estimation

The method employed in this research is adapted from existing approaches to causal modeling, specifically following the framework outlined by Rolland et al. [30]. In this approach, each variable is modeled as a function of its direct causal parents in the causal graph, along with an additive noise term. The data distribution is defined by these causal relationships, and score functions are used to identify leaf nodes within the graph. Leaf nodes are detected based on the variance of partial derivatives of the score function, which helps distinguish parent-child relationships among variables. The nodes in the graph are arranged in order by finding and removing leaf nodes one by one. The Jacobian of the score can be approximated using the Stein gradient estimator with ridge RBF kernel regression [30].

### Causal Graph Construction and Score Matching

Based on this finding provided by Rolland et al., 2022, the experiment achieves topological ordering by sequentially identifying the leaf nodes and removing them one by one. The Jacobian of the score can be approximated by Stein gradient estimator with ridge RBF kernel regression [30].

Once a topological order is estimated, the causal graph is constrained to be a subgraph of a fully connected graph. However, pruning is necessary to remove spurious edges, which is achieved using the CAM pruning process.

### CAM Pruning

The methods described above control for confounding variables by retaining key confounders during variable selection and removing irrelevant variables through correlation analysis. The CAM pruning process refines the causal graph by eliminating spurious relationships while preserving causal integrity.

After arranging the nodes, the graph is refined by using the CAM pruning process, which removes unnecessary connections to reveal the actual causal structure, aligning with methods discussed by Rolland et al.[30]. Detailed outputs are included the following metrics <sup>1</sup>:

- Structural similarity: Evaluated using SID and SHD.
- Predictive accuracy: Measured through precision, recall, and F1 score.
- Overall deviation: Assessed using L2 distance.

---

<sup>1</sup>It is noted that "Variable Selection" is to ensure that important confounders are included before pruning begins, as removing key variables early can lead to residual confounding or spurious relationships. The formal analysis of correlation removes unrelated variables-those that have no meaningful relationship with the target variable or the other variables in the system. These variables are unlikely to act as confounders since they do not introduce residual confounding or spurious relationships when removed.

For validation, after CAM pruning, the causal structure is validated using domain expertise to ensure the robustness of the inferred causal graph. CAM pruning is not a substitute for confounding control methods. It is suggested to be used in combination with other techniques to ensure the validity of causal inferences. This is also the rationale for incorporating LLMs with expertise knowledge for further exploration.

By highlighting key variables—such as access to clean fuels in rural and urban areas and urban population growth—the graph supports more targeted and effective policy-making.

## Validation: From Correlations to Causation via LLM Inquiries

In the specific context of climate change, do LLMs offer better causal inference? To address the request involving the exploration of causality factors for carbon emissions using the World Bank variables ("EG.CFT.ACCTS.RU.ZS", "EG.CFT.ACCTS.UR.ZS", "SP.URB.TOTL.IN.ZS") as the prior benchmark, the study categorizes questions into five main types for LLMs causality taxonomy prompts described in Txonomy of Causality. The study follows [24]'s CausalQuest database, with a focus on the economic and climate change context. Similarly, the study adopts Pearl's Causal Hierarchy (PCH) framework ([33, 34]), and defines a causal question as one that meets the following criteria: a question is considered causal if it involves, or if its solution process includes, any inquiry into the effects given a specific cause, and the causes given a specific effect, or the causal relationship between the given variables.

### Taxonomy of Causality

The causal taxonomy—"Direct, Preventative, Facilitative, Resultative, and Influential"—describes various types of causal relationships that verbs can imply (see SI-Table 1). This approach controls for confounding variables during LLM inquiries by leveraging a structured causal taxonomy to identify, classify, and account for different types of causal relationships[31, 32].

Table 1: Causal taxonomy used in LLM-based inquiries

Category	Description and Example
<b>Direct</b>	Refers to actions or driving forces that have an immediate impact on outcomes. The cause directly influences the effect without intermediaries. Typical verbs: "increase", "trigger" [35, 36, 37, 38]. Example: Urban access to clean fuels directly reduces carbon emissions.
<b>Preventative</b>	Causes that reduce or prevent the likelihood of a negative outcome. Verbs: "prevent", "reduce", "inhibit" [39]. Example: Improved access to clean technologies prevents an increase in carbon emissions.
<b>Facilitative</b>	Causes that enable or support an effect without directly causing it. Verbs: "enable", "allow", "support" [40, 41]. Example: Access to urban clean fuels facilitates a reduction in carbon emissions.
<b>Resultative</b>	Causes that lead to specific outcomes, emphasizing consequences. Verbs: "lead to", "result in", "cause" [42, 43]. Example: Urban population increase results in higher carbon emissions.
<b>Influential</b>	Factors that indirectly affect the likelihood or intensity of an outcome. Verbs: "influence", "impact", "affect" [44, 45, 46]. Example: Urbanization influences carbon emissions through changes in energy use patterns.