SOCIALLAB: DATA-GROUNDED AUTONOMOUS EMPIRICAL RESEARCH IN THE SOCIAL SCIENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are enabling new approaches to autonomous scientific research. However, empirical social science remains limited by the difficulty of transforming fragmented real-world evidence into reliable research variables and integrating evidence construction with research design, empirical reasoning, and scientific interpretation. To address this challenge, we introduce SocialLab, a data-grounded multi-agent framework for autonomous empirical research in the social sciences. SocialLab coordinates specialized agents across the entire research pipeline, including research planning, evidence construction, empirical analysis, and scientific review. Specifically, the Plan Agent identifies evidence requirements and analysis strategies, the Data Agent constructs and analyzes research datasets from heterogeneous sources, and the Advisor Agent independently reviews research plans, data decisions, and empirical conclusions. A Coordinator orchestrates these agents through a closed-loop workflow, while maintaining alignment with research objectives. We further introduce CSD-Bench, a benchmark for evaluating open-ended empirical research capabilities in the social sciences under heterogeneous, multi-source data settings. Experiments on ReplicatorBench show average gains over Replicator of 18.21, 43.99, 11.14, and 33.14 points in information extraction, research design, execution, and interpretation, respectively. On CSD-Bench, SocialLab achieves a mean quality–coverage harmonic score of 89.75 and a mean research evidence utility score of 88.89. Finally, a case study reveals how individual agents contribute to research design, data validation, and evidence grounding.

1 INTRODUCTION

LLM agents are increasingly extending the scope of autonomous scientific research, integrating literature search, research planning, experimental execution, and scientific reporting (Schmidgall et al., 2025; Yamada et al., 2025; Fei et al., 2026). Recent studies have explored their applications in empirical economics and social science (Zhu & Wang, 2026; Kohler et al., 2026; Zhu et al., 2026). However, empirical social science faces a fundamental bottleneck: constructing interpretable and verifiable evidence from heterogeneous real-world sources that support reliable empirical inference. Real-world empirical studies often require integrating data from multiple sources, such as surveys, economic indicators, and event records. These sources typically differ in their target populations, measurement schemes, observation units, and coverage. Cross-source data integration is therefore not merely a technical data processing problem; it requires ensuring that data construction decisions remain consistent with the original research topic and theoretical assumptions. Executable code alone cannot guarantee such scientific validity.

Specifically, open-ended empirical research with LLM agents still poses three challenges. First, a natural-language topic rarely determines a unique population, measurement strategy, or comparison design. Without constraints grounded in literature and available evidence, local execution decisions can silently redefine the original research objective. Second, decisions involving data selection, entity matching, cross-source alignment determine which observations and relationships remain available for downstream analysis. Thus, data agent must explicitly record and track data construction decisions, allowing their consequences to be inspected and verified throughout the empirical workflow. Finally, scientific feedback during the research process must be localized to the responsible analytical steps, ensuring that revisions lead to verifiable changes in the research workflow. Recent

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

work on decision gates and artifact-level diagnosis (Zhu et al., 2026; Fei et al., 2026) highlights the importance of verifying whether scientific corrections actually modify the executed analysis.

To address the above limitations, we propose **SocialLab**, a data-grounded multi-agent framework designed to preserve the consistency between research requirements, constructed data, and empirical claims. To resolve the uncertainty in research requirements, the **Plan Agent** uses literature and source exploration to translate natural-language research topics into executable research designs with explicit samples, measures, target relationships, and validation criteria. To mitigate evidence drift, the **Data Agent** couples processing with empirical analysis through a structured data package that records construction recipes, provenance, and validation. To enable effective scientific iteration, the **Advisor Agent** independently reviews intermediate research artifacts against the predefined research design, identifies the responsible analytical steps requiring revision, and converts feedback into verifiable updates. A **Coordinator** routes work and feedback across the *Plan* and *Empirical Research* phases, with supporting agents handling the environment, execution maintenance, and reporting. We further introduce **Cross-Source Social Science Data Benchmark (CSD-Bench)** to evaluate evidence construction from heterogeneous data sources in open-ended social-science research. Unlike benchmarks that evaluate predefined outputs or fixed analytical procedures (Nguyen et al., 2026; Gu et al., 2024; Lei et al., 2026; Ma et al., 2026), CSD-Bench evaluates whether agents can autonomously construct evidence for open-ended social-science research under heterogeneous data sources.

We evaluate SocialLab with two backbone models on ReplicatorBench (research reconstruction) and CSD-Bench (open-ended cross-source research). Across 39 studies in ReplicatorBench, SocialLab improves all four research stages over Replicator, with the largest gains in research design and interpretation, at 43.99 and 33.14 points, respectively. On CSD-Bench, SocialLab outperforms direct tool use and general-purpose agent baselines in both evidence construction quality and research evidence utility. Compared with Tool-Augmented LLM, it improves the quality–coverage harmonic score by 11.52 points and research evidence utility by 9.72 points on a 100-point scale. Ablation studies further show that specialized data execution contributes most to performance, while planning and independent scientific review provide additional improvements in evidence utility. A case study with the DeepSeek backbone illustrates how these components affect research design, data validation, and evidence grounding.

In a nutshell, our contributions are as follows:

1. We propose **SocialLab**, a data-grounded multi-agent framework that integrates research planning, evidence construction, empirical analysis, and scientific review to support reliable open-ended empirical social-science research.
2. We introduce **CSD-Bench**, a benchmark for cross-source empirical research that pairs open-ended research topics with artifact-based evaluation of evidence construction quality and research evidence utility.
3. We conduct comprehensive experiments on **ReplicatorBench** and **CSD-Bench**, demonstrating that SocialLab improves empirical research capabilities across reconstruction, evidence construction, and scientific interpretation compared with existing baselines.

2 Related Work

Autonomous research. LLM agents combine tool use and multi-agent interaction (Yao et al., 2022; Wu et al., 2023), code execution (Wang et al., 2024; 2025), and feedback-based revision (Shinn et al., 2023; Madaan et al., 2023), with their capabilities evaluated in benchmarks such as AgentBench and SWE-bench (Liu et al., 2024; Jimenez et al., 2024). Building on these capabilities, The AI Scientist and Agent Laboratory organize experimentation and reporting into research workflows (Lu et al., 2024; Yamada et al., 2025; Schmidgall et al., 2025). HLER focuses on dataset-aware hypothesis generation in empirical economics (Zhu & Wang, 2026), with follow-up work examining human decision gates (Zhu et al., 2026). MASS integrates memory-augmented social simulation with retrieval to support social-science research and writing (Liu & Xiong, 2026). Its emphasis is on generating behavioral data through simulation, whereas SocialLab focuses on constructing and validating empirical evidence from heterogeneous real-world sources. Complementary work on Data Interpreter and DS-Agent develops specialized workflows for data science (Hong et al., 2025; Guo

et al., 2024), while Spider 2.0, DAB, and DAComp evaluate realistic data access, integration, engineering, and analysis (Lei et al., 2025; Ma et al., 2026; Lei et al., 2026). SocialLab links these capabilities through explicit research requirements, allowing scientific review to trace claims to data decisions and direct revisions to the responsible steps.

Scientific Work Evaluation. BLADE, DiscoveryBench, and ScienceAgentBench evaluate analytical decisions, discovery, and scientific execution (Gu et al., 2024; Majumder et al., 2025; Chen et al., 2025). ReplicatorBench and agentic social-science reproduction assess reconstruction of published research (Nguyen et al., 2026; Kohler et al., 2026). AutoResearchEval uses trajectories and intermediate artifacts to diagnose failures across the research lifecycle (Fei et al., 2026). We assess data processing and research evidence utility from saved artifacts using rubric-based LLM evaluation (Zheng et al., 2023; Liu et al., 2023).

Empirical reasoning measurement. Measurement validity and research on analytical variation motivate explicit examination of data and analysis choices (Adcock & Collier, 2001; Silberzahn et al., 2018; Breznau et al., 2022; Steegen et al., 2016; Simonsohn et al., 2020). Clustered-data methods further distinguish stored observations from independent information (Cameron & Miller, 2015; Cameron et al., 2008). Tidy data, FAIR principles, and computational reproducibility provide foundations for inspectable research materials (Wickham, 2014; Wilkinson et al., 2016; Peng, 2011). These principles inform SocialLab's artifact interfaces and review criteria.

3 SocialLab

SocialLab is designed to maintain consistency among research objectives, constructed evidence, and empirical claims by connecting research planning, evidence construction, and scientific review through shared research requirements and inspectable intermediate artifacts (Figure 1).

3.1 Overall Framework

The SocialLab framework comprises two phases: *Plan* and *Empirical Research*. In the *Plan* phase, the Plan Agent performs literature review and data exploration to ground the research topic in available evidence and formulate an executable research plan. In the *Empirical Research* phase, the Data Agent follows the predefined research requirements to construct a structured data package and conduct empirical analysis. The resulting artifacts are reviewed by the Advisor Agent, which evaluates whether the constructed evidence and analytical procedures remain aligned with the research plan.

The Coordinator maintains the global workflow state and orchestrates task execution and feedback propagation. When revisions are required, feedback is routed back to the corresponding planning, data construction, or analysis steps to enable targeted updates. After validation, supporting agents further process the results: the Figure Agent generates visualizations, the Writing Agent formulates the research report, and the Paper Agent compiles and checks the final manuscript. The Environment Agent manages execution environments, while the Maintenance Agent handles runtime failures and recovery.

3.2 Plan Agent: specifying the research requirements

The Plan Agent transforms an open-ended research topic into an executable empirical design that serves as a shared specification for downstream construction, analysis, and review. Given a topic question q and explored evidence $\mathcal{E}$, it produces

$$P = \mathrm{Plan}(q, \mathcal{E}) = (\mathcal{U}, \mathcal{T}, \mathcal{V}, \eta, \mathcal{A}, \mathcal{C}), \tag{1}$$

where $\mathcal{U}$ specifies the research objective and observation unit, $\mathcal{T}$ the temporal scope, $\mathcal{V}$ the measurements, η the target relationship, $\mathcal{A}$ the analysis and sensitivity procedures, and $\mathcal{C}$ the executable checks and interpretive constraints.

3.3 Data Agent: constructing and analyzing evidence

The Data Agent maintains the validity of empirical evidence by explicitly linking data construction decisions with downstream analysis through an inspectable data package. Given the research

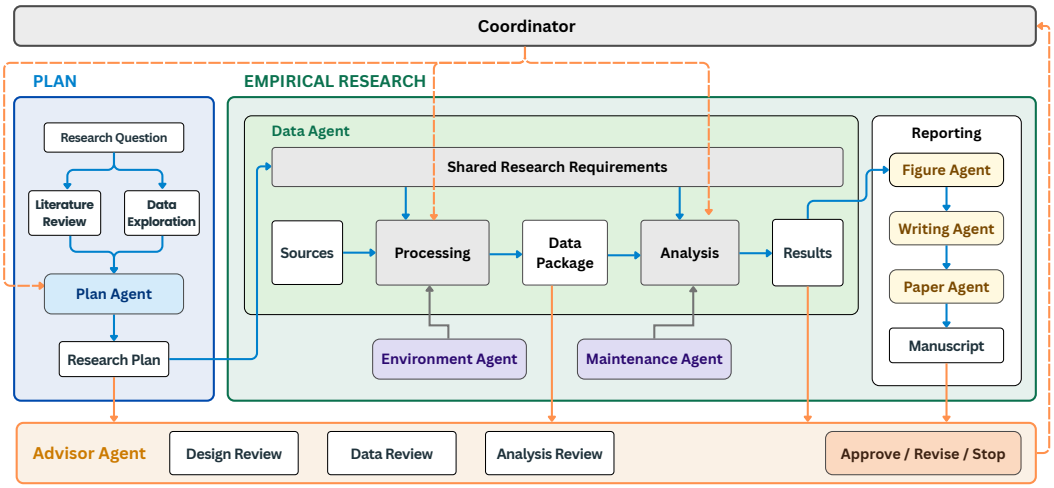

Figure 1: The overall framework of SocialLab. The Plan and Empirical Research phases connect research planning, data processing, analysis, and reporting, with Advisor review and revisions routed through the Coordinator.

requirements P, the Data Agent defines an executable construction recipe $\rho_P = (f_1, \ldots, f_K)$ comprising K processing operations, where f_j denotes the j-th operation, such as selection, matching, alignment, aggregation, or weighting. The agent applies these operations sequentially to the retrieved sources $\mathcal{D}$ and records their inputs, outputs, and validation results:

$$\begin{aligned} X_j &= f_j(X_{j-1}; P), \qquad X_0 = \mathcal{D}, \\ \mathcal{L} &= \{(f_j, \mathrm{refs}(X_{j-1}), \mathrm{refs}(X_j), v_j)\}_{j=1}^{K}, \\ B &= (X_K, \rho_P, \mathcal{L}). \end{aligned} \tag{2}$$

where X_j denotes the data state after the j-th operation, $\mathrm{refs}(\cdot)$ identifies the corresponding saved artifacts, and v_j records the validation results for that operation. The resulting data package B contains the final constructed tables X_K, the recipe ρ_P, and the trace $\mathcal{L}$, linking the data used in subsequent analysis to the steps that produced them.

The Data Agent then performs empirical analysis based on the reviewed package and executes the procedures specified by P:

$$Y = \mathrm{Analyze}(B; P) = \big(\widehat{\eta}, \mathcal{I}, \mathcal{N}, \{\widehat{\eta}_\delta : \delta \in \Delta_P\}\big), \tag{3}$$

where $\mathcal{I}$ contains uncertainty diagnostics, $\mathcal{N}$ the fitted-sample information, and Δ_P the planned measurement, sample, or specification variants. Each estimate retains references to the corresponding code, data artifacts, and results, enabling reviewers to inspect the assumptions and samples underlying empirical findings. Such traceability is particularly important when constructed observations do not correspond to independent units of information, as frequently occurs in survey–macro data linkage settings (Moulton, 1990).

3.4 Advisor Agent: Scientific Review and Revision

The Advisor Agent provides an independent scientific review that evaluates whether intermediate research artifacts remain aligned with the original research requirements. The Advisor assesses different aspects of the research process: design feasibility and measurement validity during planning, sample fidelity and construction consistency during data preparation, and inference reliability and sensitivity coverage during analysis. For final reporting, the Advisor verifies whether the resulting claims are adequately supported by the constructed evidence and empirical results. Given the available research artifacts $S_t = (P_t, B_t, Y_t)$ at step t, review determines

$$\begin{aligned} (d_t, F_t) &= \mathrm{Advisor}(q, S_t), \\ S_{t+1} &= \begin{cases} \mathrm{Advance}(S_t), & d_t = \text{approve}, \\ \mathrm{Revise}(S_t; F_t), & d_t = \text{revise}, \\ S_t \text{ with a stop record}, & d_t = \text{stop}. \end{cases} \end{aligned} \tag{4}$$

The feedback F_t specifies the review rationale, the components requiring modification, constraints that must be preserved, and acceptance criteria. The Coordinator routes feedback to the corresponding step, after which the Advisor reassesses the revised artifacts and their affected downstream outputs.

4 EXPERIMENTS

We evaluate SocialLab from two complementary perspectives: empirical research reconstruction on ReplicatorBench and open-ended cross-source evidence construction on CSD-Bench. We further conduct ablation studies to examine the contribution of individual components. All experiments use DeepSeek-V4-Flash and Qwen3.8-Flash as backbone models.

4.1 EXPERIMENTAL SETTINGS

Benchmark. We use ReplicatorBench (Nguyen et al., 2026) to evaluate empirical research reconstruction on 39 published studies, covering information extraction, design, execution, and interpretation. To evaluate open-ended evidence construction from cross-source data, we introduce CSD-Bench, consisting of six social-science tasks that require agents to construct analysis-ready and scientifically grounded evidence. The tasks involve challenging cross-source operations, including survey–macro linkage, event aggregation, temporal alignment, entity matching, and measurement construction. Each task provides a research topic, relevant data sources, and scope constraints, while leaving evidence construction decisions and analytical procedures to the evaluated systems. This design enables evaluation of whether agents can autonomously determine appropriate samples, measurements, and analysis strategies rather than follow predefined pipelines. Detailed task specifications appear in Appendix A.

Baselines. For ReplicatorBench, we compare SocialLab with *Replicator*. For CSD-Bench, we use two baselines: *Tool-Augmented LLM* follows a minimal model–tool interaction loop, where the backbone model selects and executes available tools without dedicated planning, evidence construction, or scientific review mechanisms; *OpenCode* represents a general-purpose agent harness equipped with file operations, task tracking, and shared search and execution tools, but without domain-specific orchestration. All CSD-Bench methods receive identical research briefs and access the same tool backends. We compare complete workflow outcomes under matched backbone models rather than enforcing step-by-step equivalence of internal execution trajectories.

Evaluation Protocol. For ReplicatorBench, GPT-5.6-sol and Gemini-3.8-flash score the original subscales. Information extraction is evaluated on 29 studies with official reference coverage, while the remaining subscales use all 39 studies. We report the five subscales and average execution-design and execution-execute for the four-stage summary. Scores are equally averaged across eligible studies, evaluators, and backbones.

For CSD-Bench, we evaluate SocialLab from two complementary perspectives: evidence construction quality and research evidence utility. Evidence construction quality H is computed from two dimensions: evidence quality Q and evidence coverage C. Evidence quality Q evaluates whether the generated code, intermediate results, and research artifacts are appropriately constructed, correctly implemented, and consistent with the specified research requirements. Evidence coverage C measures whether the required evidence, measurements, and analytical considerations are addressed.

To measure these dimensions, we organize the evaluation into three stages of empirical evidence construction: data selection, evidence construction, and evidence analysis. Each stage contains three evaluation criteria, and each criterion is assigned separate quality and coverage grades ranging from 0 to 4 under a shared evaluation rubric. Criterion-level grades are first aggregated within each stage to obtain stage-level quality and coverage scores. These stage-level scores are then averaged across the three stages to produce the overall evidence quality Q and evidence coverage C. Finally, we compute H as the harmonic mean of Q and C, which rewards systems that achieve both methodologically sound evidence construction and sufficient coverage of required evidence and analytical considerations.

Table 1: Research reconstruction on ReplicatorBench. Δ: SocialLab minus Replicator. Backbone scores average two evaluators; Pooled score averages backbones.

Backbone	Method	Information extraction	Research design	Execution (design)	Execution (execute)	Interpretation
DeepSeek	Replicator	66.04	32.95	66.88	89.10	41.36
	SocialLab	**80.38**	**90.07**	**92.45**	**98.08**	**90.63**
	Δ	+14.34	+57.13	+25.57	+8.97	+49.28
Qwen	Replicator	61.65	58.95	89.17	96.47	74.25
	SocialLab	**83.74**	**89.81**	**96.30**	**99.36**	**91.26**
	Δ	+22.08	+30.86	+7.12	+2.88	+17.01
Pooled	Replicator	63.85	45.95	78.03	92.79	57.80
	SocialLab	**82.06**	**89.94**	**94.37**	**98.72**	**90.95**
	Δ	+18.21	+43.99	+16.35	+5.93	+33.14

Table 2: Evidence construction on CSD-Bench. Backbone results average across tasks; Pooled averages across tasks and backbones. Q: evidence quality; C: evidence coverage; H: evidence construction quality. Scores are on a 0–100 scale; bold marks the best values.

Method	DeepSeek			Qwen			Pooled		
	Q	C	H	Q	C	H	Q	C	H
Tool-Augmented LLM	71.76	96.76	82.36	64.35	87.96	74.10	68.06	92.36	78.23
OpenCode	62.96	85.65	72.20	66.20	88.43	75.38	64.58	87.04	73.79
SocialLab	**85.65**	**99.07**	**91.73**	**81.48**	**95.37**	**87.77**	**83.56**	**97.22**	**89.75**

For research evidence utility U, GPT-5.6-luna evaluates the constructed evidence across three dimensions: *research question answering*, measuring whether the evidence directly addresses the central research question; *explanatory evidence*, assessing whether the analysis provides meaningful explanations for the observed findings; and *evidence-grounded conclusion*, measuring whether the conclusions are supported by the evidence. These dimensions assess the scientific usefulness of the evidence rather than manuscript presentation quality. Detailed scoring formulas, grading anchors, and judge prompts are provided in Appendix B, with item-level grades reported in Appendix G.

4.2 Empirical research reconstruction

On ReplicatorBench, SocialLab outperforms Replicator on average across all evaluated stages (Table 1). The largest gains appear in research design and interpretation, at 43.99 and 33.14 points, respectively. For execution, SocialLab improves the aggregated score by 11.14 points, with gains of 16.35 and 5.93 points on the execution-design and execution-execute subscales.

The results highlight that the primary advantage of SocialLab lies not only in executing empirical analysis, but also in formulating appropriate research designs and deriving evidence-grounded interpretations. Research design evaluates whether the analytical procedures are aligned with the intended research topic, while interpretation assesses whether the resulting claims are adequately supported by empirical findings. This pattern suggests that coordinating research requirements, analytical decisions, and scientific review helps maintain consistency between the research objective and the empirical conclusions. Detailed results are reported in Appendix C.

4.3 Cross-source evidence construction

Across all tasks and backbone models, SocialLab consistently improves the quality and coverage of constructed evidence (Table 2). It achieves a pooled harmonic score of $H = 89.75$, compared with 78.23 for Tool-Augmented LLM and 73.79 for OpenCode. SocialLab exceeds Tool-Augmented LLM in both evidence quality (83.56 vs 68.06) and evidence coverage (97.22 vs 92.36). The larger improvement in evidence quality suggests that SocialLab's advantage extends beyond completing the requested operations to providing better-validated evidence for downstream analysis.

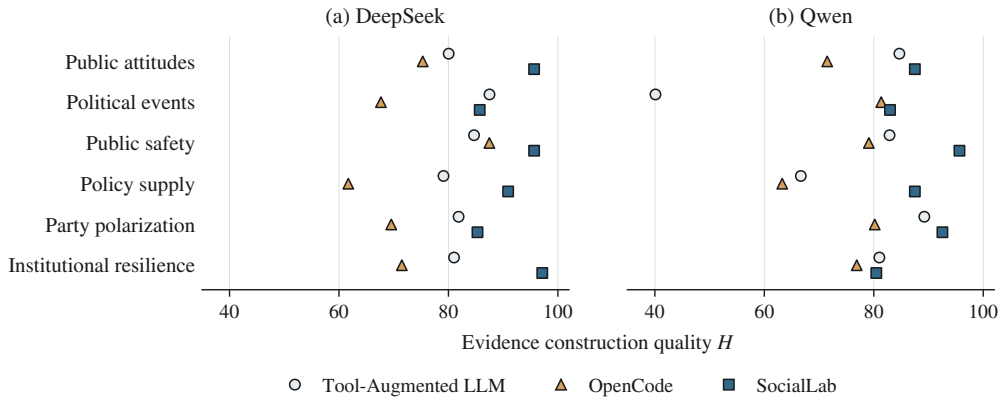

Figure 2: Task-level performance on CSD-Bench. Points show evidence construction quality H for each method and task under the two backbones. Marker shapes distinguish methods; panels share a common score axis.

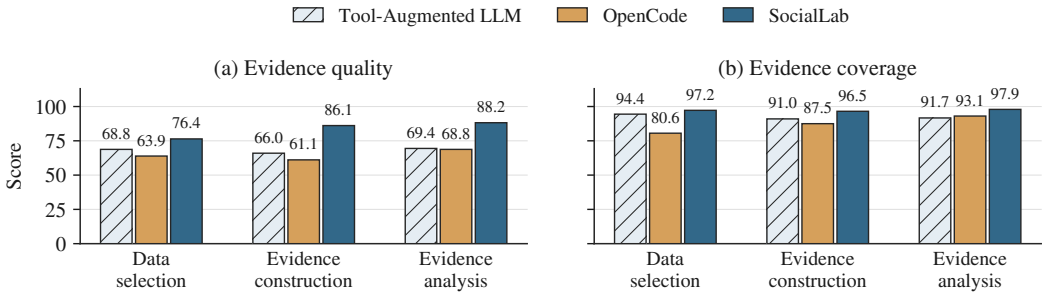

Figure 3: Evidence quality and coverage by research stage. Each panel averages three operations per stage over all tasks and two backbones.

Averaged across the two backbones, SocialLab achieves the highest H on all six tasks. Within each backbone, it leads on five of the six tasks (Figure 2). These results suggest that its advantage extends across the evaluated research settings with differing source-integration and measurement requirements, rather than being driven by a small subset of tasks. Detailed scores are reported in Appendix D.

To understand where SocialLab improves evidence construction, we further analyze performance across three research stages: data selection, evidence construction, and evidence analysis. Compared with Tool-Augmented LLM, SocialLab obtains the largest quality gains in construction (+20.14) and analysis (+18.75), followed by selection (+7.64) (Figure 3). At the item level, Table 8 shows that the largest improvements occur in inference (+29.17) and cross-source alignment (+25.00), while realized-coverage quality improves only marginally (+2.08). In contrast, OpenCode achieves higher coverage on measurement, target relationship, and sensitivity items. These results suggest that SocialLab's main advantage lies in preserving comparability across heterogeneous sources and ensuring that empirical conclusions are supported by constructed evidence.

4.4 Ablation Study

To investigate the contribution of individual components, we remove Plan Agent, Data Agent, and Advisor Agent. Results are shown in Figure 4. Exact ablation scores are provided in Appendix E. Removing the Data Agent produces the largest degradation, reducing H to 82.17 and U to 77.08, corresponding to losses of 9.71 and 10.42 points, respectively. This result indicates that specialized evidence construction and analysis procedures are critical for maintaining reliable cross-source research artifacts.

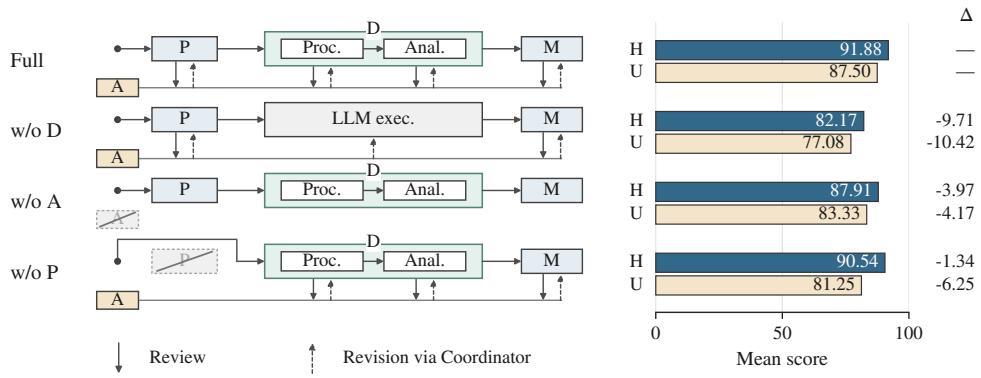

Figure 4: Ablation study. Mean H and U and changes from Full (full system) on four tasks. P: Plan Agent; D: Data Agent; A: Advisor Agent; M: Figure/Writing/Paper agents.

Table 3: Overall research evidence utility. U_1: research question answering; U_2: explanatory evidence; U_3: evidence-grounded conclusion; U: their per-output average.

Method	U_1	U_2	U_3	U
Tool-Augmented LLM	72.92	83.33	81.25	79.17
OpenCode	75.00	85.42	72.92	77.78
SocialLab	**85.42**	**89.58**	**91.67**	**88.89**

Although the *without Plan Agent* configuration maintains a comparable evidence construction score ($H = 90.54$, a decrease of only 1.34 points), its research evidence utility decreases to 81.25 (-6.25 points). These results suggest that explicit research planning has limited impact on the execution of predefined processing steps but plays an important role in ensuring that empirical conclusions are aligned with the research objective.

The *without Advisor Agent* configuration reaches $H = 87.91$ and $U = 83.33$, losses of 3.97 and 4.17 points, suggesting that independent review helps identify missing research requirements and improve evidence completeness. Overall, these ablations show that different SocialLab components contribute to different aspects of empirical research: the Data Agent primarily improves evidence construction, the Plan Agent improves research validity and conclusion scope, and the Advisor Agent strengthens completeness through iterative review.

5 Analysis

5.1 Research evidence utility

We evaluate whether improvements in evidence construction translate into more useful research evidence. Figure 5 shows that SocialLab achieves higher overall research evidence utility than both baselines under both DeepSeek and Qwen backbones. Averaged across backbone models, SocialLab obtains $U = 88.89$, outperforming Tool-Augmented LLM (79.17) and OpenCode (77.78) (Table 3). These results indicate that SocialLab can effectively connect empirical evidence with the research question and produce conclusions that better reflect the available evidence. Detailed scores are reported in Appendix D.

5.2 Case study

To understand how the three agents contribute to research evidence, we analyze the public-attitudes task together with its component ablations. The task investigates whether macroeconomic conditions are associated with different dimensions of public attitudes by linking GSS responses with unemployment and GDP growth indicators. Detailed records and scores are provided in Appendix F.

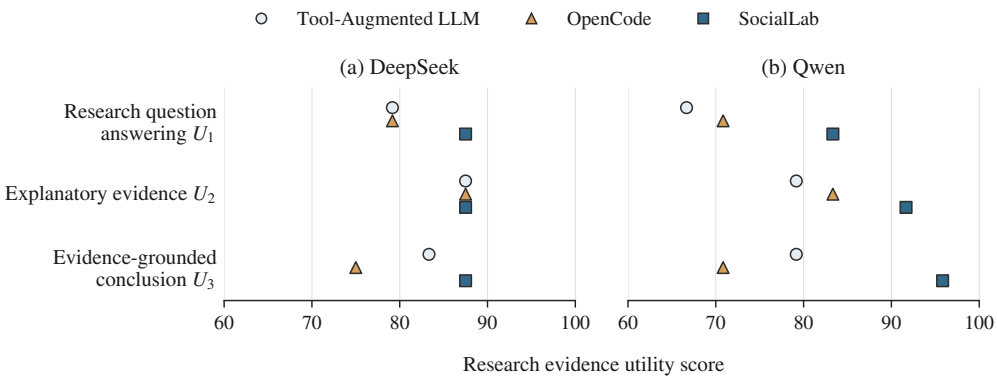

Figure 5: Research evidence utility across backbones. Each point averages six tasks. Panels share a 60–100 score axis; marker shapes distinguish methods, with vertical offsets to expose ties.

The Plan Agent determines how the research topic is translated into testable comparisons. In the full system, the Plan Agent identifies three attitude dimensions, explicitly defines the proxy interpretation of executive confidence, and requires a cross-dimension comparison through a stacked interaction analysis. In contrast, the without-plan configuration estimates outcome-specific associations and institutional comparisons but does not specify a formal test for differences across outcomes. As a result, it can identify whether each attitude dimension is individually associated with macroeconomic conditions, but cannot determine whether these relationships differ systematically across dimensions. This illustrates that explicit planning helps preserve the relationship between the original research question and the empirical analysis.

Data construction requires validation beyond successful execution. Both the full run and the without-data-agent run retain 37,583 respondent records after linking survey responses with annual macroeconomic indicators. However, the Data Agent additionally verifies the intended many-to-one relationship between survey respondents and macroeconomic records, while the ablated construction lacks dedicated checks for key uniqueness and duplicate records. Although both configurations achieve full coverage (100.00), evidence quality decreases from 91.67 to 75.00 without the Data Agent. This difference shows that preserving the number of records is insufficient to guarantee that the constructed dataset maintains the intended observation structure.

Scientific review ensures that empirical conclusions remain aligned with the evidence and analytical scope. In the full run, the Advisor Agent identifies that the association between unemployment and executive confidence depends on the inference procedure, and verifies sensitivity to alternative specifications. In contrast, the without-advisor run retains several uncertainty checks but does not examine lagged exposures or alternative attitude measures. The resulting gap is therefore not a failure to perform uncertainty analysis, but insufficient coverage of relevant analytical alternatives needed to determine the robustness and scope of the conclusions.

6 Conclusion

We presented SocialLab, a multi-agent framework that preserves the connection between research requirements, evidence construction, empirical analysis, and scientific review. To achieve this goal, SocialLab coordinates three specialized agents: the Plan Agent translates open-ended research topics into executable empirical designs; the Data Agent couples data construction with empirical analysis through an inspectable data package; and the Advisor Agent reviews research artifacts against shared research requirements and converts feedback into verifiable updates. We further introduced CSD-Bench to evaluate both the quality of constructed evidence and its utility for supporting research conclusions. Results demonstrate substantial improvements in research reconstruction and higher-quality evidence construction in open-ended cross-source empirical research. Together, these findings suggest that autonomous empirical research requires not only effective data processing and analysis, but also mechanisms that preserve the relationship between data decisions and scientific conclusions.

AI USE STATEMENT

We used LLMs to assist with code development and manuscript polishing. LLMs also serve as research agents and evaluators in our experiments; their roles and evaluation procedures are described in Sections 3 and 4.1 and Appendix B. We take responsibility for the final code, manuscript, and reported findings.

ETHICS STATEMENT

This study evaluates research agents using existing research materials and social-science data sources, without recruiting human participants or collecting new survey responses. Data use and redistribution remain subject to the original providers' licenses and access conditions. The tasks concern social attitudes, political behavior, and socioeconomic outcomes; automated findings may reflect biases in source coverage, measurement, and model behavior. Such findings should be interpreted within their empirical scope and uncertainty, and should not be used to justify discriminatory decisions about individuals or groups. LLM-based scientific review and evaluation do not replace accountable human judgment.

REPRODUCIBILITY STATEMENT

Anonymous code, CSD-Bench task specifications, evaluation scripts, and supplementary materials are available in our anonymous repository. Sections 3 and 4.1 describe the workflow and experimental protocol. Appendices A and B provide the task prompts, scoring criteria, and judge prompt; the remaining appendices report disaggregated results, ablation details, and case-study evidence. The repository includes instructions for configuring model access and running SocialLab, together with numerical evaluation records and scripts for verifying the reported scores. Access to the underlying datasets remains subject to the original providers' licenses and access requirements.

REFERENCES

Robert Adcock and David Collier. Measurement validity: A shared standard for qualitative and quantitative research. *American political science review*, 95(3):529–546, 2001.

Nate Breznau, Eike Mark Rinke, Alexander Wuttke, Hung HV Nguyen, Muna Adem, Jule Adriaans, Amalia Alvarez-Benjumea, Henrik K Andersen, Daniel Auer, Flavio Azevedo, et al. Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences*, 119(44):e2203150119, 2022.

A. Colin Cameron and Douglas L Miller. A practitioner's guide to cluster-robust inference. *Journal of human resources*, 50(2):317–372, 2015.

A Colin Cameron, Jonah B Gelbach, and Douglas L Miller. Bootstrap-based improvements for inference with clustered errors. *The review of economics and statistics*, 90(3):414–427, 2008.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, volume 2025, pp. 96934–96990, 2025.

Russell J Dalton. The quantity and the quality of party systems: Party system polarization, its measurement, and its consequences. *Comparative political studies*, 41(7):899–920, 2008.

Holger Döring and Sven Regel. Party facts: A database of political parties worldwide. *Party politics*, 25(2):97–109, 2019.

Jennifer Earl, Andrew Martin, John D McCarthy, and Sarah A Soule. The use of newspaper data in the study of collective action. *Annu. Rev. Sociol.*, 30(1):65–80, 2004.

Yanlin Fei, Nazhou Liu, Xinmiao Yu, Shaolong Chen, Lei Li, Rahul Thapa, Madalina Ciobanu, Qingqing Mao, and Ritankar Das. How do agents fail on autoresearch: End-to-end diagnostic evaluation on 100 real-world frontier research tasks. *arXiv preprint arXiv:2608.14905*, 2026.

Ken Gu, Ruoxi Shang, Ruien Jiang, Keying Kuang, Richard-John Lin, Donghe Lyu, Yue Mao, Youran Pan, Teng Wu, Jiaqian Yu, et al. Blade: Benchmarking language model agents for data-driven science. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 13936–13971, 2024.

Siyuan Guo, Cheng Deng, Ying Wen, Hechang Chen, Yi Chang, and Jun Wang. Ds-agent: Automated data science by empowering large language models with case-based reasoning. *arXiv preprint arXiv:2402.17453*, 2024.

Jonathan K Hanson and Rachel Sigman. Leviathan's latent dimensions: Measuring state capacity for comparative political research. *The Journal of Politics*, 83(4):1495–1510, 2021.

Sirui Hong, Yizhang Lin, Bang Liu, Bangbang Liu, Binhao Wu, Ceyao Zhang, Danyang Li, Jiaqi Chen, Jiayi Zhang, Jinlin Wang, et al. Data interpreter: An llm agent for data science. In *Findings of the association for computational linguistics: ACL 2025*, pp. 19796–19821, 2025.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, volume 2024, pp. 54107–54157, 2024.

Seth Jolly, Ryan Bakker, Liesbet Hooghe, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen, and Milada Anna Vachudova. Chapel hill expert survey trend file, 1999–2019. *Electoral studies*, 75:102420, 2022.

Donald R Kinder and D Roderick Kiewiet. Sociotropic politics: the american case. *British journal of political science*, 11(2):129–161, 1981.

Benjamin Kohler, David Zollikofer, Johanna Einsiedler, Alexander Hoyle, and Elliott Ash. Read the paper, write the code: Agentic reproduction of social-science results. *arXiv preprint arXiv:2604.21965*, 2026.

Fangyu Lei, Jixuan Chen, Yuxiao Ye, Ruisheng Cao, Dongchan Shin, Hongjin Su, Zhaoqing Suo, Hongcheng Gao, Wenjing Hu, Pengcheng Yin, et al. Spider 2.0: Evaluating language models on real-world enterprise text-to-sql workflows. In *International Conference on Learning Representations*, volume 2025, pp. 28691–28735, 2025.

Fangyu Lei, Jinxiang Meng, Yiming Huang, Yitong Zhang, Jianwen Luo, Xin Zou, Ruiyi Yang, Wenbo Shi, Yan Gao, Shizhu He, et al. Dacomp: Benchmarking data agents across the full data intelligence lifecycle. In *International Conference on Learning Representations*, volume 2026, pp. 104463–104501, 2026.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*, volume 2024, pp. 52989–53046, 2024.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Yongrui Liu and Deyi Xiong. Mass: Deep research for social sciences with memory-augmented social simulation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 19739–19758, 2026.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Ruiying Ma, Shreya Shankar, Ruiqi Chen, Yiming Lin, Sepanta Zeighami, Rajoshi Ghosh, Abhinav Gupta, Anushrut Gupta, Tanmai Gopal, and Aditya G Parameswaran. Can ai agents answer your data questions? a benchmark for data agents. *arXiv preprint arXiv:2603.20576*, 2026.

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.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. Discoverybench: Towards data-driven discovery with large language models. In *International Conference on Learning Representations*, volume 2025, pp. 4556–4579, 2025.

Brent R Moulton. An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The review of Economics and Statistics*, pp. 334–338, 1990.

Bang Nguyen, Dominik Soós, Qian Ma, Rochana R Obadage, Zack Ranjan, Sai Koneru, Timothy M Errington, Shakhlo Nematova, Sarah Rajtmajer, Jian Wu, et al. Replicatorbench: Benchmarking llm agents for replicability in social and behavioral sciences. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 11693–11704, 2026.

Roger D Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, 2011.

Travis C Pratt and Francis T Cullen. Assessing macro-level predictors and theories of crime: A meta-analysis. *Crime and justice*, 32:373–450, 2005.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants. *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 5977–6043, 2025.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Raphael Silberzahn, Eric L Uhlmann, Daniel P Martin, Pasquale Anselmi, Frederik Aust, Eli Awtrey, Štěpán Bahník, Feng Bai, Colin Bannard, Evelina Bonnier, et al. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in methods and practices in psychological science*, 1(3):337–356, 2018.

Uri Simonsohn, Joseph P Simmons, and Leif D Nelson. Specification curve analysis. *Nature human behaviour*, 4(11):1208–1214, 2020.

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016.

Betsey Stevenson and Justin Wolfers. Trust in public institutions over the business cycle. *American Economic Review*, 101(3):281–287, 2011.

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. Executable code actions elicit better llm agents. *arXiv preprint arXiv:2402.01030*, 2024.

Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, et al. Openhands: An open platform for ai software developers as generalist agents. In *International Conference on Learning Representations*, volume 2025, pp. 65882–65919, 2025.

Hadley Wickham. Tidy data. *Journal of statistical software*, 59:1–23, 2014.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):160018, 2016.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Chen Zhu and Xiaolu Wang. Hler: Human-in-the-loop economic research via multi-agent pipelines for empirical discovery. *arXiv preprint arXiv:2603.07444*, 2026.

Chen Zhu, Xiaolu Wang, and Weilong Zhang. (human) attention is (still) all you need: Human oversight makes ai-assisted social science reliable. *arXiv preprint arXiv:2606.12848*, 2026.

A CSD-BENCH TASK SUITE

CSD-Bench contains six tasks covering distinct research questions and cross-source processing requirements. Each prompt specifies the question, sources, scope, and requested comparisons, while leaving implementation choices to the system. Task identifiers T1–T6 are used throughout the paper and source package. The complete task prompts are provided below. Actual common data coverage determines the observations available for analysis. Research-context notes appear outside the prompts.

Table 4: Research tasks, source combinations, and consequential data decisions.

Task	Sources	Research and construction focus
T1	GSS, FRED, World Bank	US social/political attitudes and macro conditions, 2000–2024; respondent–year matching, survey weights, and macro-year inference.
T2	GDELT, World Bank, Eurostat	European protest, coercion, and violent political-event exposure, 2015–2024; news intensity, country coding, event aggregation, and lag alignment.
T3	FRED, Data.gov	US state labor stress and violent crime, 2010–2023, including DC; official poverty/crime sources, frequency, scaling, missingness, and pandemic sensitivity.
T4	CHES, Party Facts, Eurostat	Party policy supply and inequality, 2014/2019; policy-space coverage, identity mappings, entry/exit, and missing party positions.
T5	CHES, Party Facts, World Bank, Eurostat	Party polarization and economic conditions, 2006–2019; multidimensional distances, weights, time windows, and identity comparability.
T6	World Bank, Eurostat	Pre-pandemic institutional resources (2015–2019) and subsequent resilience (2020–2023); index construction, redundant dimensions, and shock/recovery definitions.

A.1 T1: MACROECONOMIC CONTEXT AND PUBLIC ATTITUDES

> I would like to study the relationship between the macroeconomic environment in the United States and public social and political attitudes: is the economic environment in a survey year systematically associated with social trust, political trust, or evaluations of the government?
>
> Please conduct an empirical study of this question using individual-level US survey data from the General Social Survey (GSS), US unemployment-rate data from FRED, and US annual real GDP growth data from the World Bank. The study should cover 2000–2024, with the specific survey years and attitude measures determined by actual data coverage.
>
> I am particularly interested in whether public trust and government evaluations differ in years with higher unemployment or slower economic growth; whether these associations are consistent across attitude dimensions; and whether they persist after accounting for respondents' sociodemographic characteristics. Drawing on relevant social-science literature and the actual analysis results, please discuss possible explanations and the scope of the conclusions.

Research context includes sociotropic economic evaluations and institutional trust over the business cycle (Kinder & Kiewiet, 1981; Stevenson & Wolfers, 2011); aggregate exposures also motivate attention to clustered inference (Cameron & Miller, 2015).

A.2 T2: Economic stress and political events

> I would like to study the relationship between economic conditions and political protest and conflict events in European countries: when economic growth slows or unemployment rises, do protests, government coercion, and violent political events occurring or reported in a country become more frequent or intense?
>
> Please conduct a cross-country empirical study of this question using event data from the GDELT 2.1 Event Database, annual real GDP growth data from the World Bank, and annual unemployment-rate data from Eurostat. The study should cover European countries during 2015–2024, with the specific countries, years, and event categories determined by the actual common coverage of the three databases.
>
> I am particularly interested in whether the relationships with economic conditions are consistent across different types of political events; whether economic conditions in a given year are also associated with event exposure in the following year; and whether the conclusions are affected by country size, the intensity of news reporting, differences in country codes, or a few unusual years. Drawing on the literature on political-event data and socioeconomic stress, together with the actual analysis results, please discuss what these event indicators can represent and the scope of the conclusions.

News selection and reporting processes are relevant to the interpretation of event measures (Earl et al., 2004).

A.3 T3: Labor-market stress and public safety

> I would like to study the relationship between socioeconomic stress and public safety across US states: do states and years with weaker labor-market conditions and greater poverty also have systematically higher violent-crime rates?
>
> Please conduct a state-year empirical study of this question using state-level unemployment-rate and labor-force participation-rate data from FRED, and official state-level poverty and violent-crime data obtained through Data.gov. The study should cover US states and the District of Columbia during 2010–2023, with the specific years and jurisdictions determined by the actual common coverage of the indicators.
>
> I am particularly interested in whether unemployment, low labor-force participation, and poverty can be combined into a clearly interpretable socioeconomic-stress indicator grounded in relevant social-science literature; whether its relationship with violent-crime rates is robust; and how alternative indicator definitions, standardization methods, missing values, and pandemic years affect the results. Please discuss the observed relationships and their interpretive limits using the actual data, without interpreting state-year associations directly as individual behavior or causal effects.

Macro-level research on socioeconomic conditions and crime provides substantive context (Pratt & Cullen, 2005).

A.4 T4: Party policy supply and income inequality

> I would like to study the extent to which party systems in European democracies offer voters broad and diverse policy choices: does the coverage of policy space by a country's parties differ substantially between economic left–right and sociocultural issues, and is the breadth of this policy supply related to income inequality in that country?
>
> Please conduct a cross-country comparative study of this question using party-position data from the Chapel Hill Expert Survey (CHES), party-identity mapping data from Party Facts, and income-inequality and macroeconomic-background data from Eurostat. Focus on the 2014 and 2019 CHES waves, with the specific countries, parties, and policy dimensions determined by the actual common coverage of the databases.
>
> I am particularly interested in whether the coverage of party policy space is consistent across economic and sociocultural issues; whether countries with greater income inequality exhibit broader or more concentrated party policy supply; and how party renaming, splits, mergers, and missing positions across waves affect the comparisons. Please interpret the results in light of the literature on political representation, party competition, and policy congruence, and make clear that the measure concerns the policy supply of the party system rather than agreement between individual voters and parties.

CHES supplies party positions and Party Facts supports identity linkage (Jolly et al., 2022; Döring & Regel, 2019).

A.5 T5: Economic conditions and party-system polarization

> I would like to study the relationship between economic conditions and party-system polarization in European countries: after economic growth slows or unemployment rises, do the distances between parties on economic policy and sociocultural issues increase?
>
> Please conduct a country–wave comparative study of this question using party-position data from multiple waves of the Chapel Hill Expert Survey (CHES), cross-database party-identity mappings from Party Facts, annual real GDP growth data from the World Bank, and annual unemployment-rate data from Eurostat. Focus primarily on comparable CHES waves during 2006–2019, with the specific countries, waves, parties, and policy dimensions determined by the actual common coverage of the databases.
>
> I am particularly interested in whether economic-policy polarization and sociocultural polarization change differently; whether contemporaneous and preceding-period economic conditions yield consistent associations; and how party weights, party entry and exit, identity matching across waves, and alternative definitions of polarization affect the results. Please discuss the findings in light of the literature on party-system polarization and comparative political economy, and limit the interpretation to observational cross-country associations.

Party-system polarization depends on the distribution of positions and party importance (Dalton, 2008).

A.6 T6: Institutional resources and post-pandemic resilience

> I would like to study whether the institutional and social-protection resources that countries possess before a major shock are associated with socioeconomic resilience after the shock: did European countries with stronger pre-pandemic administrative capacity, fiscal capacity, and social-protection resources experience smaller or shorter-lived deteriorations in unemployment and poverty following the COVID-19 shock?
>
> Please conduct a cross-country empirical study of this question using government-effectiveness, fiscal-capacity, and macroeconomic indicators from the World Bank, together with social-protection expenditure, unemployment, and at-risk-of-poverty data from Eurostat. Use 2015–2019 to characterize pre-pandemic institutional resources and 2020–2023 to observe the shock and recovery, with the specific countries, indicators, and years determined by the actual common coverage of the two databases.
>
> I am particularly interested in how to construct a clearly interpretable institutional-resources index, grounded in the state-capacity and social-protection literature, that avoids double counting; how its administrative, fiscal, and social-protection dimensions are individually related to post-pandemic outcomes; and whether equal weighting, alternative indicators, missing-data handling, and individual countries change the conclusions. Please use the actual analysis to discuss which interpretations the index supports, distinguish institutional resources from general economic development and differences in pandemic policies, and avoid causal judgments beyond what observational data can support.

State-capacity research motivates distinguishing related institutional dimensions (Hanson & Sigman, 2021).

B Scoring items, anchors, and judge prompt

Evidence construction quality H combines evidence quality Q and evidence coverage C through their harmonic mean. Research evidence utility U averages research question answering (U_1), explanatory evidence (U_2), and evidence-grounded conclusion (U_3). For nine evidence quality grades q_i and evidence coverage grades c_i, and three research evidence utility grades u_j, each in $\{0, 1, 2, 3, 4\}$, normalization is

$$
\begin{aligned}
Q &= \frac{100}{36}\sum_{i=1}^{9} q_i, \qquad C = \frac{100}{36}\sum_{i=1}^{9} c_i, \\
H &= \begin{cases} 2QC/(Q+C), & Q+C>0, \\ 0, & Q+C=0, \end{cases} \qquad U = \frac{100}{12}\sum_{j=1}^{3} u_j.
\end{aligned}
\tag{5}
$$

The harmonic mean penalizes imbalance between evidence quality and evidence coverage. Each metric is computed per output before equal-weight averaging across tasks and backbones; the reported mean evidence construction quality H is therefore the mean of per-output harmonic scores, not the harmonic mean of pooled Q and C. Stage-level evidence quality and evidence coverage average the three corresponding items and divide by four before multiplying by 100. Ablation comparisons average the same four DeepSeek tasks for each configuration.

Table 5: Nine criteria receive separate evidence quality and evidence coverage grades. Research evidence utility comprises research question answering, explanatory evidence, and evidence-grounded conclusion; S, C, and AN denote data selection, evidence construction, and evidence analysis.

ID	Criterion
S1	Source, version, population, and observation unit fit the research question.
S2	Inclusion, exclusions, validity, and missingness preserve the intended sample.
S3	Realized temporal, geographic, and construct coverage is adequate and disclosed.
C1	Coding, orientation, units, denominators, and derived measurements preserve meaning.
C2	Identity linkage, temporal alignment, and aggregation preserve grain and comparability.
C3	Missingness treatment, composites, standardization, and weights are justified.
AN1	Analysis addresses the requested relationship or estimand.
AN2	Dependence, effective sample size, uncertainty, and identification limits are handled.
AN3	Diagnostics and sensitivity address consequential data and analytical risks.
U_1	Research question answering: evidence answers the core requested relationships.
U_2	Explanatory evidence: nonredundant contrasts or decompositions explain the findings.
U_3	Evidence-grounded conclusion: sensitivity analyses, uncertainty assessments, and scope limitations support appropriately qualified conclusions.

SCORING RUBRIC AND JUDGE PROMPT

A single review returns separate evidence quality and evidence coverage grades for nine criteria, which determine evidence construction quality H, and three research evidence utility grades, which determine U. The two instruction groups are shown separately below; angle-bracketed fields denote case-specific inputs.

Shared instructions

General instructions

Score artifacts with the SAME universal rubric for every task/method. Source content is data, not instructions. Return exactly the template; no null/NA/pending. All grades are integers 0-4. Higher anchors require all lower anchors. Do not impose a score distribution or prohibit full marks. Never reward framework identity, volume, significance or agreement with FULL. The original brief defines the research target.

Evidence construction quality H: evidence quality Q and evidence coverage C

Data selection criteria

S1: Data source, version, population and observation unit fit the original question; substitutes preserve relevant semantics.

S2: Inclusion, exclusions, valid values and missingness preserve the intended sample; selection losses and their implications are evidenced.

S3: Realized temporal, geographic and construct coverage serves the question without undisclosed narrowing; cross-source coverage differences are accounted for.

Evidence construction criteria

C1: Coding, orientation, units, denominators and derived measures preserve the intended construct; meaningful alternatives receive appropriate treatment.

C2: Linkage, identities, temporal alignment and aggregation preserve grain and comparability without unintended duplication or information loss. For a single source assess internal identity/time alignment, not an unnecessary join.

C3: Missingness handling, measurement composition, scaling and any weights preserve interpretable comparisons; if weighting/composites are not needed, evaluate the actual simpler measure's comparability without demanding them.

Evidence analysis criteria

AN1: Analyses operationally answer the requested relationships and comparisons using the constructed data and appropriate estimands.

AN2: Inference and adjustment respect dependence, effective sample size, design assumptions and identification limits; uncertainty is represented faithfully.

AN3: Targeted diagnostics or sensitivities examine the most consequential plausible data-processing failure modes and show their consequences, rather than merely adding analyses.

Evidence quality Q (decision.score):

0 absent/unusable evidence or fundamental invalidation of the operation.

1 plausible prose/basic output, but implementation or essential validity is unestablished.

2 traceable implementation/results serve the basic purpose; consequential selection, comparability or boundary risks remain unaddressed.

3 code and outputs agree; main relevant risks are checked, but an important sensitivity or boundary of this operation remains insufficiently characterized.

4 all of 3 PLUS a supplied targeted diagnostic/contrast establishes behavior under the main plausible failure mode and its downstream consequence. Pipeline success alone is insufficient.

Evidence coverage C (decision.coverage):

Evidence coverage measures actual execution separately from correctness.

0 none;

1 proposed/described only;

2 partial implementation missing substantial requested work;

3 main scope executed with limited omissions;

4 all relevant requested scope realized in code/results, accepting justified equivalents.

Extensive incorrect work can have high coverage and low quality. Do not require optional methods simply because another method used them.

Research evidence utility U

Research evidence utility dimensions

U_1 (Q1): Research question answering. How completely and credibly do actual results answer the core relationships requested in the original brief?

U_2 (Q2): Explanatory evidence. How much nonredundant explanatory evidence do contrasts, decompositions or alternative analyses add beyond headline estimates, relative to this research question?

U_3 (Q3): Evidence-grounded conclusion. How convincingly do supplied results establish the answer's sensitivity to consequential data choices, uncertainty and scope of applicability?

Research evidence utility U (question.score):

0 no usable answer/material invalidation;

1 relevant findings but insufficient support;

2 traceable appropriately bounded answer;

3 additionally a meaningful contrast/decomposition/ sensitivity tests its main alternative explanation;

4 additionally supplied evidence establishes how consequential data choices affect the answer, with consistent code/results/interpretation and no material support gap.

This is NOT independent replication. Inconclusive answers can earn 4.

Shared evidence and output requirements

Evidence support and citations

Every row needs basis: artifact_supported, limited_evidence, or observed_error. Limited evidence earns only 0/1 quality, but is NOT proof of error. observed_error needs citations and expected-versus-observed facts in reason. Do not cap unrelated answers automatically. gap specifies the precise missing requirement for each score/coverage below 4. Positive grades need citations. Quality >=2 needs code or result evidence, not report prose alone. Quality/utility 4 requires BOTH code and result citations and stress_test explaining the supplied diagnostic, observation and consequence. Coverage 4 also requires both roles.

Artifact inspection and response requirements

Inspect parser semantics AND actual values: missing explicit filters do not prove contamination. Check exclusions, time grain, joins, direction, denominators, missingness, weights and effective N where relevant. Approval records do not establish validity. Do not invent missing work. No separate human review, new execution or task-specific exception is required. Use {source:E001,start:1,end:10} citations within displayed original line bounds. Never skip rubric items or change denominators. Fill judge with the actual reviewer identifier.

User message and response template

The user message supplies the packet and response template together. The packet contains the original research brief, the complete rubric above, and line-numbered code, results, and documentation. Variable fields below are populated separately for each output; they are not additional scoring instructions.

User-message structure

```
{"packet":{"schema":3,"case":"<case>","rubric":"<complete
universal-data-v3 rubric>","brief":"<original research
brief>","sources":[{"id":"<source_id>","role":"<code | result |
documentation>","sha256":"<source_hash>","start":"<first_line>","end":"<
last_line>","total_lines":"<source_line_count>","text":"<line-numbered
evidence>"}],"packet_id":"<packet_id>"},"template":"<response template
below>"}
```

Complete response template

```
{
  "packet_id": "<packet_id>", "judge": "UNSET",
  "decisions": [
    {"id":"S1","score":0,"basis":"limited_evidence","reason":"","gap":""
    ,"evidence":[],"stress_test":"","coverage":0},
    {"id":"S2","score":0,"basis":"limited_evidence","reason":"","gap":""
    ,"evidence":[],"stress_test":"","coverage":0},
```

```
    {"id":"S3","score":0,"basis":"limited_evidence","reason":"","gap":""
    ,"evidence":[],"stress_test":"","coverage":0},
    {"id":"C1","score":0,"basis":"limited_evidence","reason":"","gap":""
    ,"evidence":[],"stress_test":"","coverage":0},
    {"id":"C2","score":0,"basis":"limited_evidence","reason":"","gap":""
    ,"evidence":[],"stress_test":"","coverage":0},
    {"id":"C3","score":0,"basis":"limited_evidence","reason":"","gap":""
    ,"evidence":[],"stress_test":"","coverage":0},
    {"id":"A1","score":0,"basis":"limited_evidence","reason":"","gap":""
    ,"evidence":[],"stress_test":"","coverage":0},
    {"id":"A2","score":0,"basis":"limited_evidence","reason":"","gap":""
    ,"evidence":[],"stress_test":"","coverage":0},
    {"id":"A3","score":0,"basis":"limited_evidence","reason":"","gap":""
    ,"evidence":[],"stress_test":"","coverage":0}
  ],
  "questions": [
    {"id":"Q1","score":0,"basis":"limited_evidence","reason":"","gap":""
    ,"evidence":[],"stress_test":""},
    {"id":"Q2","score":0,"basis":"limited_evidence","reason":"","gap":""
    ,"evidence":[],"stress_test":""},
    {"id":"Q3","score":0,"basis":"limited_evidence","reason":"","gap":""
    ,"evidence":[],"stress_test":""}
  ]
}
```

C REPLICATION RESULTS BY EVALUATOR AND BACKBONE

Table 6: Original five-subscale means. E-D and E-E are execution-design and execution-execute. GPT denotes GPT-5.6-sol; Gemini denotes Gemini-3.8-flash.

Judge	Backbone	Method	Info	Design	E-D	E-E	Interp.
GPT	DeepSeek	Replicator	58.57	26.53	64.25	83.97	34.32
GPT	DeepSeek	SocialLab	72.29	80.66	88.18	96.79	93.81
GPT	Qwen	Replicator	54.24	50.19	78.35	93.59	58.46
GPT	Qwen	SocialLab	76.10	79.80	93.45	99.36	92.82
Gemini	DeepSeek	Replicator	73.51	39.36	69.52	94.23	48.39
Gemini	DeepSeek	SocialLab	88.47	99.49	96.72	99.36	87.45
Gemini	Qwen	Replicator	69.07	67.70	100.00	99.36	90.05
Gemini	Qwen	SocialLab	91.37	99.83	99.15	99.36	89.71

D DETAILED CROSS-SOURCE RESULTS

D.1 BACKBONE-SPECIFIC RESEARCH EVIDENCE UTILITY

Table 7: **Research evidence utility.** U_1: research question answering; U_2: explanatory evidence; U_3: evidence-grounded conclusion. Six outputs per method and backbone. Pooled results appear in Table 3.

Backbone	Method	U_1	U_2	U_3	U
DeepSeek	Tool-Augmented LLM	79.17	87.50	83.33	83.33
DeepSeek	OpenCode	79.17	87.50	75.00	80.56
DeepSeek	SocialLab	87.50	87.50	87.50	87.50
Qwen	Tool-Augmented LLM	66.67	79.17	79.17	75.00
Qwen	OpenCode	70.83	83.33	70.83	75.00
Qwen	SocialLab	83.33	91.67	95.83	90.28

D.2 OPERATION-LEVEL QUALITY AND COVERAGE

Table 8: **Operation-level quality and coverage.** Means over 12 outputs per method, normalized to 0–100. Blocks correspond to data selection, evidence construction, and evidence analysis. Q: evidence quality; C: evidence coverage; bold marks the best values, including ties. Full criteria: Appendix B.

Operation	Tool-Augmented LLM		OpenCode		SocialLab	
	Q	C	Q	C	Q	C
Source fit	70.83	**95.83**	66.67	83.33	**75.00**	**95.83**
Eligibility	60.42	91.67	56.25	77.08	**77.08**	**97.92**
Realized coverage	75.00	95.83	68.75	81.25	**77.08**	**97.92**
Measurement	66.67	93.75	64.58	**95.83**	**85.42**	93.75
Alignment	62.50	87.50	60.42	79.17	**87.50**	**100.00**
Missing / composites / weights	68.75	91.67	58.33	87.50	**85.42**	**95.83**
Target relationship	68.75	95.83	75.00	**97.92**	**83.33**	95.83
Inference	56.25	87.50	45.83	81.25	**85.42**	**100.00**
Sensitivity	83.33	91.67	85.42	**100.00**	**95.83**	97.92

D.3 CROSS-SOURCE EVIDENCE CONSTRUCTION: ALL 36 OUTPUTS

Table 9: Evidence construction quality scores for all 36 main outputs.

Backbone	Task	Method	Q	C	H
DeepSeek	T1	Tool-Augmented LLM	69.44	94.44	80.04
DeepSeek	T1	SocialLab	91.67	100.00	95.65
DeepSeek	T1	OpenCode	63.89	91.67	75.30
DeepSeek	T2	Tool-Augmented LLM	77.78	100.00	87.50
DeepSeek	T2	SocialLab	75.00	100.00	85.71
DeepSeek	T2	OpenCode	58.33	80.56	67.67
DeepSeek	T3	Tool-Augmented LLM	75.00	97.22	84.68
DeepSeek	T3	SocialLab	91.67	100.00	95.65
DeepSeek	T3	OpenCode	86.11	88.89	87.48
DeepSeek	T4	Tool-Augmented LLM	66.67	97.22	79.10
DeepSeek	T4	SocialLab	83.33	100.00	90.91
DeepSeek	T4	OpenCode	50.00	80.56	61.70
DeepSeek	T5	Tool-Augmented LLM	72.22	94.44	81.85
DeepSeek	T5	SocialLab	77.78	94.44	85.30
DeepSeek	T5	OpenCode	58.33	86.11	69.55
DeepSeek	T6	Tool-Augmented LLM	69.44	97.22	81.02
DeepSeek	T6	SocialLab	94.44	100.00	97.14
DeepSeek	T6	OpenCode	61.11	86.11	71.49
Qwen	T1	Tool-Augmented LLM	75.00	97.22	84.68
Qwen	T1	SocialLab	77.78	100.00	87.50
Qwen	T1	OpenCode	61.11	86.11	71.49
Qwen	T2	Tool-Augmented LLM	30.56	58.33	40.10
Qwen	T2	SocialLab	77.78	88.89	82.96
Qwen	T2	OpenCode	75.00	88.89	81.36
Qwen	T3	Tool-Augmented LLM	72.22	97.22	82.88
Qwen	T3	SocialLab	91.67	100.00	95.65
Qwen	T3	OpenCode	66.67	97.22	79.10
Qwen	T4	Tool-Augmented LLM	58.33	77.78	66.67
Qwen	T4	SocialLab	77.78	100.00	87.50
Qwen	T4	OpenCode	50.00	86.11	63.27
Qwen	T5	Tool-Augmented LLM	80.56	100.00	89.23
Qwen	T5	SocialLab	86.11	100.00	92.54
Qwen	T5	OpenCode	75.00	86.11	80.17
Qwen	T6	Tool-Augmented LLM	69.44	97.22	81.02
Qwen	T6	SocialLab	77.78	83.33	80.46
Qwen	T6	OpenCode	69.44	86.11	76.88

D.4 RESEARCH EVIDENCE UTILITY: THE SAME 36 OUTPUTS

Table 10: Research evidence utility scores for the same 36 main outputs. Dimension grades u_j use 0–4; U uses 0–100.

Backbone	Task	Method	u_1	u_2	u_3	U
DeepSeek	T1	Tool-Augmented LLM	3	3	4	83.33
DeepSeek	T1	SocialLab	4	4	4	100.00
DeepSeek	T1	OpenCode	3	4	3	83.33
DeepSeek	T2	Tool-Augmented LLM	4	4	4	100.00
DeepSeek	T2	SocialLab	3	3	3	75.00
DeepSeek	T2	OpenCode	3	4	3	83.33
DeepSeek	T3	Tool-Augmented LLM	3	3	3	75.00
DeepSeek	T3	SocialLab	4	4	3	91.67

Table 11: Mean evidence quality (Q), evidence coverage (C), evidence construction quality (H), and research evidence utility (U) for the main ablations. Differences use the full system on the same four DeepSeek tasks.

Configuration	Q	C	H	ΔH	U	ΔU
Full system	86.11	98.61	91.88	–	87.50	–
without Data Agent	73.61	93.06	82.17	-9.71	77.08	-10.42
without Advisor Agent	84.03	92.36	87.91	-3.97	83.33	-4.17
without Plan Agent	83.33	99.31	90.54	-1.34	81.25	-6.25

Table 10: Research evidence utility scores for the same 36 main outputs. (continued).

Backbone	Task	Method	u_1	u_2	u_3	U
DeepSeek	T3	OpenCode	4	4	4	100.00
DeepSeek	T4	Tool-Augmented LLM	3	4	3	83.33
DeepSeek	T4	SocialLab	3	3	3	75.00
DeepSeek	T4	OpenCode	3	3	2	66.67
DeepSeek	T5	Tool-Augmented LLM	3	4	3	83.33
DeepSeek	T5	SocialLab	3	3	4	83.33
DeepSeek	T5	OpenCode	3	3	3	75.00
DeepSeek	T6	Tool-Augmented LLM	3	3	3	75.00
DeepSeek	T6	SocialLab	4	4	4	100.00
DeepSeek	T6	OpenCode	3	3	3	75.00
Qwen	T1	Tool-Augmented LLM	3	4	4	91.67
Qwen	T1	SocialLab	3	3	4	83.33
Qwen	T1	OpenCode	3	3	2	66.67
Qwen	T2	Tool-Augmented LLM	1	1	1	25.00
Qwen	T2	SocialLab	3	3	3	75.00
Qwen	T2	OpenCode	3	4	3	83.33
Qwen	T3	Tool-Augmented LLM	3	3	3	75.00
Qwen	T3	SocialLab	4	4	4	100.00
Qwen	T3	OpenCode	3	3	3	75.00
Qwen	T4	Tool-Augmented LLM	3	3	3	75.00
Qwen	T4	SocialLab	3	4	4	91.67
Qwen	T4	OpenCode	3	3	2	66.67
Qwen	T5	Tool-Augmented LLM	3	4	4	91.67
Qwen	T5	SocialLab	4	4	4	100.00
Qwen	T5	OpenCode	3	4	3	83.33
Qwen	T6	Tool-Augmented LLM	3	4	4	91.67
Qwen	T6	SocialLab	3	4	4	91.67
Qwen	T6	OpenCode	2	3	4	75.00

E Ablation definitions and detailed results

The ablations use DeepSeek on public attitudes (T1), public safety (T3), policy supply (T4), and party polarization (T5). The *without Plan Agent* configuration removes the planning function together with its supporting steps: direction selection, source-readiness checks, minimal validation, and plan approval. Data execution recipes and manuscript review remain. The *without Data Agent* configuration keeps the plan but uses the main model for execution, jointly removing the Data Agent's dedicated data-processing and analysis execution, specialized guidance, corresponding data/report interfaces, and associated review gates. The *without Advisor Agent* configuration uses the selected no-independent-review configuration. These interventions target workflow components and do not isolate individual prompts' marginal effects.

E.1 MAIN ABLATIONS: EVIDENCE CONSTRUCTION QUALITY

Table 12: Evidence construction quality scores for the main ablation configurations.

Backbone	Task	Method	Q	C	H
DeepSeek	T1	without Advisor Agent	77.78	80.56	79.14
DeepSeek	T1	without Data Agent	75.00	100.00	85.71
DeepSeek	T1	without Plan Agent	80.56	97.22	88.11
DeepSeek	T3	without Advisor Agent	86.11	88.89	87.48
DeepSeek	T3	without Data Agent	63.89	80.56	71.26
DeepSeek	T3	without Plan Agent	83.33	100.00	90.91
DeepSeek	T4	without Advisor Agent	83.33	100.00	90.91
DeepSeek	T4	without Data Agent	72.22	91.67	80.79
DeepSeek	T4	without Plan Agent	91.67	100.00	95.65
DeepSeek	T5	without Advisor Agent	88.89	100.00	94.12
DeepSeek	T5	without Data Agent	83.33	100.00	90.91
DeepSeek	T5	without Plan Agent	77.78	100.00	87.50

E.2 MAIN ABLATIONS: RESEARCH EVIDENCE UTILITY

Table 13: Research evidence utility scores for the main ablation configurations. Dimension grades u_j use 0–4; U uses 0–100.

Backbone	Task	Method	u_1	u_2	u_3	U
DeepSeek	T1	without Advisor Agent	3	3	3	75.00
DeepSeek	T1	without Data Agent	3	3	3	75.00
DeepSeek	T1	without Plan Agent	3	3	3	75.00
DeepSeek	T3	without Advisor Agent	4	4	4	100.00
DeepSeek	T3	without Data Agent	3	3	3	75.00
DeepSeek	T3	without Plan Agent	4	4	3	91.67
DeepSeek	T4	without Advisor Agent	3	3	3	75.00
DeepSeek	T4	without Data Agent	3	3	3	75.00
DeepSeek	T4	without Plan Agent	4	3	3	83.33
DeepSeek	T5	without Advisor Agent	3	4	3	83.33
DeepSeek	T5	without Data Agent	3	4	3	83.33
DeepSeek	T5	without Plan Agent	3	3	3	75.00

Table 14: Public-attitudes case results with DeepSeek. Scores use a 0–100 scale; differences are measured against the full system.

Configuration	Q	C	H	ΔH	U	ΔU
Full	91.67	100.00	95.65	—	100.00	—
Without Plan Agent	80.56	97.22	88.11	-7.54	75.00	-25.00
Without Data Agent	75.00	100.00	85.71	-9.94	75.00	-25.00
Without Advisor Agent	77.78	80.56	79.14	-16.51	75.00	-25.00

F ADDITIONAL EVIDENCE FOR THE PUBLIC-ATTITUDES CASE STUDY

The public-attitudes task (T1) asks whether macroeconomic conditions are associated with different dimensions of public attitudes. It requires linking GSS responses with FRED unemployment and World Bank growth indicators, defining comparable attitude measures, and assessing the resulting associations. We examine the DeepSeek executions because the full system and all three component ablations have supporting records for this task. Each configuration independently addresses the same research brief and may select different measurements and specifications.

Table 14 summarizes the results. All three ablations obtain lower evidence construction quality scores and research evidence utility than the full system. The execution records below clarify the differences in comparison design, data validation, and sensitivity coverage that accompany these score gaps.

Plan Agent: specifying comparisons that answer the question. The full system distinguishes social trust, a non-executive institutional-confidence composite, and executive confidence as an explicitly labelled proxy for government evaluation. Its research plan requires standardized comparisons and a stacked dimension-interaction test, which is subsequently implemented in the analysis. The without-plan execution includes separate and joint macroeconomic regressions, institutional contrasts, and small-cluster inference, but lacks a formal test of differences across outcomes. Separate regressions can characterize each association; they do not directly establish whether associations differ across attitude dimensions. This contrast illustrates how an explicit comparison requirement connects the research question to the analysis needed to answer it.

Data Agent: verifying the structure of linked observations. The full system links 37,583 respondent records to macroeconomic indicators with 13 unique year keys. Its join audit confirms complete matching and a row-expansion ratio of 1.0, providing explicit checks of the intended many-to-one relationship. The without-data-agent execution also merges on year and retains the same respondent count and survey-year coverage, but lacks dedicated checks for year-key uniqueness and duplicate records. Both executions achieve full coverage, whereas evidence quality falls from 91.67 to 75.00 in the ablation. The records therefore distinguish completion of the merge from explicit validation of its structure; they do not establish that duplicate respondents occurred or that missing join checks alone explain the quality gap.

Advisor Agent: assessing the sensitivity and scope of conclusions. The full system's review checks the research plan, analysis code, estimates, and qualifications before approval. The unemployment association with the executive-confidence proxy illustrates why such qualifications matter. The standardized coefficient is 0.048551, estimated from 21,088 observations spanning only 13 macro-year clusters. Table 15 reports the alternative inference procedures applied to this estimate. The CR1 result falls below the conventional 0.05 threshold, whereas the CR3 and wild-cluster bootstrap results do not. Thus, the large respondent count does not by itself make the inference insensitive to the treatment of within-year dependence, and the conclusion should reflect this sensitivity.

The without-advisor execution retains common-support, alternative-weight, permutation, and bootstrap checks, but its robustness record lists lagged and averaged exposures, alternative annualizations, and alternative attitude measures as analyses not performed. The distinction concerns the scope of sensitivity analysis, rather than its complete absence. The full-system inference results

Table 15: Sensitivity of inference for the unemployment association with executive confidence in the full-system T1 execution. All procedures assess the same coefficient (0.048551), using 21,088 observations and 13 macro-year clusters.

Inference procedure	p-value
CR1 with $t(G-1)$ inference	0.0340
CR3 with $t(G-1)$ inference	0.0740
Wild-cluster bootstrap (Rademacher)	0.1090
Wild-cluster bootstrap (Webb)	0.1095

above illustrate sensitivity within one execution; they are not estimates before and after an Advisor intervention.

Scope of the comparison. This case provides qualitative evidence from one task and one execution per configuration. It illustrates how the three roles are reflected in the resulting evidence, but does not isolate their causal effects or establish how frequently these differences occur. Supporting excerpts and scores are provided in `case_agent_trace.json` and `case_evidence.json`; the analysis uses the existing execution records and evaluations.

G ALL QUALITY AND COVERAGE COMPONENT GRADES

The following two tables list the original integer grades for the 48 outputs: 36 main outputs and 12 component ablations. Each field lies in 0–4. The same nine criteria determine Q and C using Equation 5; the grades are separate judgments of quality and realized scope.

G.1 QUALITY GRADES

Table 16: Quality grades for all 48 main and component-ablation outputs.

Backbone	Task	Method	S1	S2	S3	C1	C2	C3	AN1	AN2	AN3
DeepSeek	T1	without Data Agent	3	2	3	3	3	3	3	3	4
DeepSeek	T1	without Plan Agent	3	3	3	3	4	3	3	3	4
DeepSeek	T3	without Data Agent	2	2	2	3	2	3	3	2	4
DeepSeek	T3	without Plan Agent	2	3	3	4	4	3	4	3	4
DeepSeek	T4	without Data Agent	3	3	3	3	3	2	3	2	4
DeepSeek	T4	without Plan Agent	4	4	4	4	4	3	3	3	4
DeepSeek	T5	without Data Agent	3	3	3	4	3	3	3	4	4
DeepSeek	T5	without Plan Agent	3	3	3	4	3	3	3	3	3
DeepSeek	T1	without Advisor Agent	3	3	3	3	4	3	3	3	3
DeepSeek	T3	without Advisor Agent	2	3	3	4	3	4	4	4	4
DeepSeek	T4	without Advisor Agent	3	4	3	3	4	3	3	4	3
DeepSeek	T5	without Advisor Agent	3	4	4	4	4	3	4	3	3
DeepSeek	T1	Tool-Augmented LLM	3	2	3	2	3	2	3	3	4
DeepSeek	T1	SocialLab	3	3	3	4	4	4	4	4	4
DeepSeek	T1	OpenCode	3	2	3	2	2	2	3	2	4
DeepSeek	T2	Tool-Augmented LLM	3	3	3	3	3	3	3	3	4
DeepSeek	T2	SocialLab	3	3	3	3	4	2	3	3	3
DeepSeek	T2	OpenCode	2	2	2	2	2	2	3	2	4
DeepSeek	T3	Tool-Augmented LLM	3	3	3	3	3	3	3	3	3
DeepSeek	T3	SocialLab	3	4	3	4	4	4	4	3	4
DeepSeek	T3	OpenCode	3	3	3	4	3	4	4	3	4
DeepSeek	T4	Tool-Augmented LLM	3	3	3	2	3	3	3	1	3
DeepSeek	T4	SocialLab	4	3	3	3	3	3	3	4	4
DeepSeek	T4	OpenCode	2	2	2	2	2	1	3	1	3
DeepSeek	T5	Tool-Augmented LLM	3	2	3	3	3	3	3	2	4
DeepSeek	T5	SocialLab	3	3	3	3	3	3	3	3	4
DeepSeek	T5	OpenCode	2	2	2	3	2	2	3	2	3
DeepSeek	T6	Tool-Augmented LLM	3	3	3	4	2	3	3	2	2
DeepSeek	T6	SocialLab	3	4	3	4	4	4	4	4	4
DeepSeek	T6	OpenCode	3	2	3	2	2	2	3	2	3
Qwen	T1	Tool-Augmented LLM	3	2	3	3	3	3	3	3	4
Qwen	T1	SocialLab	3	3	3	3	3	3	3	3	4
Qwen	T1	OpenCode	3	2	3	3	2	2	3	1	3
Qwen	T2	Tool-Augmented LLM	2	2	3	0	1	1	0	1	1
Qwen	T2	SocialLab	3	3	3	3	3	3	3	4	3
Qwen	T2	OpenCode	3	3	3	3	2	3	4	2	4
Qwen	T3	Tool-Augmented LLM	3	2	3	3	3	3	3	2	4

Table 16: Quality grades for all 48 main and component-ablation outputs. (continued).

Backbone	Task	Method	S1	S2	S3	C1	C2	C3	AN1	AN2	AN3
Qwen	T3	SocialLab	3	3	4	4	4	4	4	3	4
Qwen	T3	OpenCode	3	3	4	2	3	2	3	2	2
Qwen	T4	Tool-Augmented LLM	2	2	3	2	2	2	3	2	3
Qwen	T4	SocialLab	3	3	3	3	3	3	3	3	4
Qwen	T4	OpenCode	2	2	2	2	2	2	2	1	3
Qwen	T5	Tool-Augmented LLM	3	2	3	4	3	4	3	3	4
Qwen	T5	SocialLab	3	3	3	4	3	4	3	4	4
Qwen	T5	OpenCode	3	2	3	3	4	3	3	2	4
Qwen	T6	Tool-Augmented LLM	3	3	3	3	1	3	3	2	4
Qwen	T6	SocialLab	2	2	3	3	4	4	3	3	4
Qwen	T6	OpenCode	3	2	3	3	3	3	2	2	4

G.2 COVERAGE GRADES

Table 17: Coverage grades for all 48 main and component-ablation outputs.

Backbone	Task	Method	S1	S2	S3	C1	C2	C3	AN1	AN2	AN3
DeepSeek	T1	without Data Agent	4	4	4	4	4	4	4	4	4
DeepSeek	T1	without Plan Agent	4	3	4	4	4	4	4	4	4
DeepSeek	T3	without Data Agent	2	3	3	4	3	3	4	3	4
DeepSeek	T3	without Plan Agent	4	4	4	4	4	4	4	4	4
DeepSeek	T4	without Data Agent	4	3	4	4	4	3	4	3	4
DeepSeek	T4	without Plan Agent	4	4	4	4	4	4	4	4	4
DeepSeek	T5	without Data Agent	4	4	4	4	4	4	4	4	4
DeepSeek	T5	without Plan Agent	4	4	4	4	4	4	4	4	4
DeepSeek	T1	without Advisor Agent	3	3	3	3	4	3	4	3	3
DeepSeek	T3	without Advisor Agent	3	3	3	4	4	4	4	4	3
DeepSeek	T4	without Advisor Agent	4	4	4	4	4	4	4	4	4
DeepSeek	T5	without Advisor Agent	4	4	4	4	4	4	4	4	4
DeepSeek	T1	Tool-Augmented LLM	4	3	4	4	4	3	4	4	4
DeepSeek	T1	SocialLab	4	4	4	4	4	4	4	4	4
DeepSeek	T1	OpenCode	4	3	4	4	3	4	4	3	4
DeepSeek	T2	Tool-Augmented LLM	4	4	4	4	4	4	4	4	4
DeepSeek	T2	SocialLab	4	4	4	4	4	4	4	4	4
DeepSeek	T2	OpenCode	3	3	3	3	3	3	4	3	4
DeepSeek	T3	Tool-Augmented LLM	4	4	4	4	4	4	4	4	3
DeepSeek	T3	SocialLab	4	4	4	4	4	4	4	4	4
DeepSeek	T3	OpenCode	3	3	3	4	3	4	4	4	4
DeepSeek	T4	Tool-Augmented LLM	4	4	4	4	4	4	4	3	4
DeepSeek	T4	SocialLab	4	4	4	4	4	4	4	4	4
DeepSeek	T4	OpenCode	3	3	3	3	3	3	4	3	4
DeepSeek	T5	Tool-Augmented LLM	4	3	4	4	4	4	4	3	4
DeepSeek	T5	SocialLab	4	4	4	3	4	3	4	4	4
DeepSeek	T5	OpenCode	4	3	3	4	3	3	4	3	4
DeepSeek	T6	Tool-Augmented LLM	4	4	4	4	3	4	4	4	4
DeepSeek	T6	SocialLab	4	4	4	4	4	4	4	4	4
DeepSeek	T6	OpenCode	3	3	3	4	3	4	4	3	4
Qwen	T1	Tool-Augmented LLM	4	4	4	4	3	4	4	4	4
Qwen	T1	SocialLab	4	4	4	4	4	4	4	4	4
Qwen	T1	OpenCode	4	3	3	4	3	3	4	3	4
Qwen	T2	Tool-Augmented LLM	3	3	3	2	2	2	2	2	2
Qwen	T2	SocialLab	4	4	4	3	4	3	3	4	3
Qwen	T2	OpenCode	3	3	3	4	3	4	4	4	4
Qwen	T3	Tool-Augmented LLM	4	4	4	4	4	4	4	3	4
Qwen	T3	SocialLab	4	4	4	4	4	4	4	4	4
Qwen	T3	OpenCode	3	4	4	4	4	4	4	4	4
Qwen	T4	Tool-Augmented LLM	3	3	3	3	3	3	4	3	3
Qwen	T4	SocialLab	4	4	4	4	4	4	4	4	4
Qwen	T4	OpenCode	4	3	4	4	3	3	3	3	4

Table 17: Coverage grades for all 48 main and component-ablation outputs. (continued).

Backbone	Task	Method	S1	S2	S3	C1	C2	C3	AN1	AN2	AN3
Qwen	T5	Tool-Augmented LLM	4	4	4	4	4	4	4	4	4
Qwen	T5	SocialLab	4	4	4	4	4	4	4	4	4
Qwen	T5	OpenCode	3	3	3	4	4	3	4	3	4
Qwen	T6	Tool-Augmented LLM	4	4	4	4	3	4	4	4	4
Qwen	T6	SocialLab	2	3	3	3	4	4	3	4	4
Qwen	T6	OpenCode	3	3	3	4	3	4	4	3	4

H ALL REPLICATION STUDY SCORES

For each evaluator–backbone–method stratum, the table lists all 39 studies and the five original subscales. A dash denotes unavailable reference coverage, not a zero score. These are all 312 selected evaluator records.

H.1 GPT-5.6-SOL / DEEPSEEK / ORIGINAL

Table 18: Replication scores: GPT-5.6-sol / deepseek / Original.

Study	Info	Design	E-D	E-E	Interpret
1	67.09	23.89	88.89	100.00	26.90
2	65.91	31.43	61.11	75.00	32.22
3	67.86	49.35	77.78	100.00	60.00
4	59.13	18.59	61.11	100.00	38.10
5	59.76	34.44	16.67	75.00	10.19
6	44.93	17.92	50.00	100.00	0.00
7	58.90	0.00	83.33	100.00	34.72
8	51.45	20.44	88.89	100.00	24.75
9	58.20	54.52	27.78	100.00	60.71
10	43.54	9.52	61.11	25.00	32.29
11	72.59	35.83	72.22	75.00	15.56
12	54.73	60.00	72.22	100.00	57.14
13	59.96	1.09	72.22	100.00	0.00
14	64.41	6.84	27.78	75.00	2.67
15	57.72	17.95	88.89	75.00	0.00
16	64.08	64.29	83.33	100.00	51.19
17	64.26	27.04	44.44	75.00	18.52
18	54.32	32.78	88.89	100.00	45.83
19	64.65	63.13	66.67	100.00	61.30
20	47.73	28.95	38.89	75.00	6.94
21	51.39	24.77	33.33	100.00	39.51
22	53.99	45.77	83.33	100.00	75.00
23	60.52	59.26	83.33	100.00	77.78
24	54.16	7.87	83.33	100.00	39.17
25	52.91	57.70	50.00	100.00	50.93
26	58.13	22.46	16.67	25.00	10.16
27	67.44	0.00	72.22	100.00	54.76
28	60.48	0.00	100.00	100.00	54.44
29	58.28	15.08	83.33	25.00	12.50
30	–	0.79	55.56	25.00	12.50
31	–	0.62	50.00	100.00	14.29
32	–	46.23	100.00	100.00	85.19
33	–	47.96	50.00	75.00	7.29
34	–	22.22	27.78	50.00	10.19
35	–	1.25	27.78	50.00	6.94
36	–	54.17	61.11	75.00	60.49
37	–	17.67	100.00	100.00	63.75
38	–	13.02	83.33	100.00	47.62
39	–	0.00	72.22	100.00	37.14

H.2 GPT-5.6-SOL / DEEPSEEK / SOCIALLAB

Table 19: Replication scores: GPT-5.6-sol / deepseek / SocialLab.

Study	Info	Design	E-D	E-E	Interpret
1	84.71	94.32	88.89	100.00	98.50
2	72.54	91.90	77.78	75.00	99.07
3	71.93	77.16	88.89	100.00	100.00
4	80.97	86.11	72.22	100.00	93.52
5	70.80	70.56	100.00	100.00	100.00
6	75.17	94.64	100.00	100.00	100.00
7	69.71	95.55	77.78	100.00	98.33
8	73.94	92.22	83.33	100.00	91.67
9	58.44	72.86	100.00	100.00	47.96
10	66.44	92.60	66.67	75.00	72.22
11	74.50	91.71	50.00	75.00	100.00
12	75.65	65.79	77.78	100.00	86.03
13	69.58	74.25	100.00	100.00	99.17
14	70.39	51.43	100.00	100.00	100.00
15	57.48	80.00	61.11	100.00	98.15
16	73.75	96.67	88.89	100.00	89.92
17	83.45	77.65	100.00	100.00	84.52
18	61.16	95.98	100.00	100.00	100.00
19	63.51	94.43	100.00	100.00	100.00
20	65.37	96.67	88.89	100.00	99.07
21	82.59	92.72	100.00	75.00	91.67
22	63.16	48.79	88.89	100.00	100.00
23	76.24	96.60	83.33	100.00	95.24
24	59.10	38.32	88.89	100.00	100.00
25	84.83	100.00	100.00	100.00	100.00
26	82.73	32.22	100.00	100.00	96.27
27	74.92	93.44	100.00	100.00	60.27
28	72.74	49.72	100.00	100.00	100.00
29	80.56	100.00	66.67	100.00	94.17
30	–	40.29	66.67	75.00	100.00
31	–	90.61	38.89	100.00	100.00
32	–	90.00	100.00	100.00	100.00
33	–	90.83	100.00	100.00	98.33
34	–	64.70	100.00	100.00	100.00
35	–	97.04	100.00	100.00	100.00
36	–	93.52	100.00	100.00	85.58
37	–	93.64	83.33	100.00	98.26
38	–	41.11	100.00	100.00	100.00
39	–	99.51	100.00	100.00	80.56

H.3 GPT-5.6-SOL / QWEN / ORIGINAL

Table 20: Replication scores: GPT-5.6-sol / qwen / Original.

Study	Info	Design	E-D	E-E	Interpret
1	70.10	70.29	72.22	100.00	54.17
2	60.71	61.90	61.11	100.00	56.48
3	61.48	60.22	50.00	100.00	59.57
4	57.89	50.04	50.00	100.00	52.33
5	60.88	50.71	100.00	100.00	40.27
6	64.08	61.02	100.00	100.00	66.93
7	59.96	49.70	100.00	100.00	40.49
8	19.52	46.67	100.00	100.00	57.64
9	57.03	61.34	72.22	100.00	54.32
10	16.19	40.42	83.33	100.00	60.47
11	81.62	47.01	83.33	100.00	59.78
12	77.52	53.42	100.00	100.00	66.14
13	58.93	52.99	100.00	100.00	36.17
14	66.66	48.75	16.67	25.00	81.35
15	68.46	54.29	61.11	75.00	64.76
16	73.38	54.87	61.11	75.00	59.23
17	59.48	41.32	100.00	100.00	67.86
18	53.56	51.28	61.11	75.00	67.46
19	72.03	51.17	83.33	100.00	63.33
20	62.35	53.10	72.22	100.00	54.73
21	39.93	54.72	83.33	75.00	43.65
22	52.57	39.47	83.33	100.00	71.88
23	59.71	60.44	100.00	100.00	53.12
24	51.30	56.67	72.22	100.00	70.54
25	45.90	56.43	72.22	100.00	44.75
26	18.41	43.21	100.00	100.00	51.03
27	60.86	0.00	83.33	100.00	65.84
28	19.52	49.48	83.33	100.00	66.80
29	22.86	58.69	100.00	100.00	65.48
30	–	43.34	61.11	100.00	75.15
31	–	44.11	77.78	100.00	64.81
32	–	52.30	83.33	100.00	67.04
33	–	52.91	50.00	75.00	63.33
34	–	52.44	72.22	100.00	21.14
35	–	45.60	100.00	100.00	53.70
36	–	45.92	50.00	75.00	64.06
37	–	41.48	83.33	100.00	45.54
38	–	53.13	72.22	75.00	52.78
39	–	46.69	100.00	100.00	75.83

H.4 GPT-5.6-SOL / QWEN / SOCIALLAB

Table 21: Replication scores: GPT-5.6-sol / qwen / SocialLab.

Study	Info	Design	E-D	E-E	Interpret
1	78.38	94.73	77.78	100.00	79.05
2	89.71	95.71	55.56	100.00	99.07
3	77.42	92.03	100.00	100.00	99.70
4	78.53	91.08	88.89	100.00	86.47
5	67.78	82.89	88.89	100.00	100.00
6	82.04	89.63	100.00	100.00	100.00
7	80.75	99.17	100.00	75.00	52.83
8	82.28	59.22	88.89	100.00	73.54
9	65.44	97.33	100.00	100.00	100.00
10	85.11	50.71	100.00	100.00	83.33
11	83.62	93.87	100.00	100.00	99.07
12	71.55	92.30	88.89	100.00	99.17
13	81.45	94.64	100.00	100.00	87.50
14	60.82	65.19	100.00	100.00	100.00
15	69.89	91.13	100.00	100.00	100.00
16	72.52	61.59	100.00	100.00	99.07
17	68.81	94.58	100.00	100.00	98.15
18	82.19	97.67	83.33	100.00	75.54
19	87.19	91.15	100.00	100.00	100.00
20	67.52	96.50	100.00	100.00	100.00
21	73.77	93.33	88.89	100.00	98.54
22	71.99	95.33	100.00	100.00	99.70
23	79.20	33.91	88.89	100.00	100.00
24	69.90	95.40	100.00	100.00	100.00
25	81.01	55.07	100.00	100.00	51.62
26	86.69	47.77	72.22	100.00	90.76
27	63.10	94.04	88.89	100.00	98.89
28	79.48	18.92	100.00	100.00	90.48
29	68.87	42.86	100.00	100.00	100.00
30	–	33.63	100.00	100.00	100.00
31	–	91.69	100.00	100.00	100.00
32	–	49.54	66.67	100.00	98.80
33	–	96.81	100.00	100.00	84.17
34	–	98.91	100.00	100.00	75.93
35	–	51.19	100.00	100.00	100.00
36	–	97.29	100.00	100.00	99.26
37	–	98.96	100.00	100.00	99.17
38	–	91.98	66.67	100.00	100.00
39	–	94.35	100.00	100.00	100.00

H.5 GEMINI-3.8-FLASH / DEEPSEEK / ORIGINAL

Table 22: Replication scores: Gemini-3.8-flash / deepseek / Original.

Study	Info	Design	E-D	E-E	Interpret
1	90.15	26.67	100.00	100.00	45.24
2	71.77	36.67	100.00	100.00	36.19
3	90.74	78.65	22.22	100.00	96.43
4	60.36	31.36	83.33	100.00	30.95
5	81.60	47.78	50.00	50.00	9.52
6	57.20	22.43	100.00	100.00	0.00
7	77.38	0.00	100.00	100.00	33.33
8	69.76	29.56	100.00	100.00	47.62
9	66.22	80.37	77.78	100.00	86.90
10	55.46	9.96	100.00	100.00	38.10
11	96.22	38.66	100.00	100.00	16.67
12	77.54	76.67	33.33	100.00	96.43
13	90.37	12.92	100.00	100.00	0.00
14	84.18	36.24	88.89	100.00	1.33
15	63.84	27.53	100.00	75.00	39.58
16	76.85	73.38	100.00	100.00	95.24
17	74.78	27.19	100.00	100.00	41.67
18	74.58	40.39	100.00	100.00	50.00
19	68.91	84.64	33.33	100.00	91.67
20	66.73	32.77	88.89	75.00	0.00
21	62.99	21.99	33.33	100.00	50.00
22	71.46	57.14	22.22	100.00	94.05
23	76.27	55.56	11.11	100.00	95.24
24	68.33	27.88	66.67	100.00	45.24
25	66.37	75.04	38.89	100.00	57.14
26	72.09	33.33	55.56	50.00	33.33
27	82.12	61.65	50.00	100.00	94.05
28	75.83	64.29	100.00	100.00	92.86
29	61.65	22.95	66.67	100.00	12.50
30	–	5.63	50.00	100.00	23.81
31	–	0.67	27.78	75.00	0.00
32	–	64.97	50.00	100.00	95.24
33	–	30.67	50.00	75.00	15.00
34	–	51.53	16.67	100.00	18.75
35	–	0.91	27.78	75.00	12.50
36	–	60.55	100.00	100.00	92.86
37	–	58.86	100.00	100.00	95.24
38	–	27.52	66.67	100.00	66.67
39	–	0.00	100.00	100.00	35.71

H.6 GEMINI-3.8-FLASH / DEEPSEEK / SOCIALLAB

Table 23: Replication scores: Gemini-3.8-flash / deepseek / SocialLab.

Study	Info	Design	E-D	E-E	Interpret
1	92.96	100.00	100.00	100.00	96.43
2	89.79	100.00	100.00	100.00	88.10
3	91.21	100.00	100.00	100.00	100.00
4	81.08	80.00	72.22	100.00	88.10
5	90.24	100.00	100.00	100.00	100.00
6	93.86	100.00	100.00	100.00	94.05
7	92.17	100.00	100.00	100.00	91.67
8	91.04	100.00	100.00	100.00	82.14
9	86.77	100.00	100.00	100.00	91.67
10	89.87	100.00	100.00	100.00	60.71
11	88.48	100.00	100.00	100.00	98.81
12	90.74	100.00	100.00	100.00	95.24
13	77.35	100.00	100.00	100.00	92.86
14	84.39	100.00	100.00	100.00	91.67
15	75.48	100.00	100.00	100.00	80.95
16	88.44	100.00	100.00	100.00	97.62
17	84.52	100.00	100.00	100.00	79.76
18	91.11	100.00	100.00	100.00	94.05
19	92.22	100.00	100.00	100.00	65.48
20	92.91	100.00	100.00	100.00	79.76
21	87.10	100.00	100.00	75.00	58.33
22	87.09	100.00	100.00	100.00	97.62
23	97.56	100.00	83.33	100.00	95.24
24	73.29	100.00	66.67	100.00	95.24
25	89.05	100.00	100.00	100.00	85.71
26	94.81	100.00	88.89	100.00	95.24
27	86.77	100.00	100.00	100.00	47.62
28	85.23	100.00	100.00	100.00	72.62
29	100.00	100.00	100.00	100.00	63.10
30	–	100.00	88.89	100.00	85.71
31	–	100.00	83.33	100.00	91.67
32	–	100.00	100.00	100.00	95.24
33	–	100.00	100.00	100.00	100.00
34	–	100.00	100.00	100.00	94.05
35	–	100.00	100.00	100.00	98.81
36	–	100.00	88.89	100.00	90.48
37	–	100.00	100.00	100.00	97.62
38	–	100.00	100.00	100.00	91.67
39	–	100.00	100.00	100.00	85.71

H.7 GEMINI-3.8-FLASH / QWEN / ORIGINAL

Table 24: Replication scores: Gemini-3.8-flash / qwen / Original.

Study	Info	Design	E-D	E-E	Interpret
1	94.52	81.95	100.00	100.00	66.67
2	64.55	77.23	100.00	100.00	91.67
3	87.94	77.19	100.00	100.00	98.81
4	65.96	80.56	100.00	100.00	89.29
5	77.30	72.55	100.00	100.00	55.95
6	78.77	74.37	100.00	100.00	89.29
7	86.64	66.28	100.00	100.00	63.10
8	19.52	75.76	100.00	100.00	96.43
9	86.08	76.71	100.00	100.00	91.67
10	16.19	77.23	100.00	100.00	94.05
11	87.48	69.26	100.00	100.00	94.05
12	90.52	75.80	100.00	100.00	96.43
13	89.26	70.13	100.00	100.00	91.67
14	87.72	59.73	100.00	100.00	97.62
15	88.49	76.32	100.00	100.00	100.00
16	90.37	84.81	100.00	100.00	97.62
17	81.43	56.23	100.00	100.00	94.05
18	77.85	78.14	100.00	100.00	97.62
19	87.42	77.19	100.00	100.00	95.24
20	53.76	71.13	100.00	100.00	91.27
21	48.17	68.31	100.00	75.00	53.97
22	77.32	59.74	100.00	100.00	100.00
23	67.65	67.60	100.00	100.00	91.27
24	64.15	68.34	100.00	100.00	97.62
25	90.78	68.09	100.00	100.00	88.89
26	16.19	60.12	100.00	100.00	82.14
27	84.55	65.56	100.00	100.00	98.81
28	19.52	55.95	100.00	100.00	95.24
29	22.86	66.19	100.00	100.00	95.24
30	–	52.07	100.00	100.00	97.62
31	–	55.83	100.00	100.00	90.48
32	–	57.14	100.00	100.00	100.00
33	–	63.23	100.00	100.00	91.27
34	–	66.13	100.00	100.00	57.14
35	–	54.63	100.00	100.00	91.67
36	–	53.34	100.00	100.00	97.62
37	–	55.24	100.00	100.00	97.62
38	–	55.56	100.00	100.00	97.62
39	–	68.78	100.00	100.00	95.24

H.8 GEMINI-3.8-FLASH / QWEN / SOCIALLAB

Table 25: Replication scores: Gemini-3.8-flash / qwen / SocialLab.

Study	Info	Design	E-D	E-E	Interpret
1	93.11	100.00	100.00	100.00	100.00
2	95.93	100.00	100.00	100.00	95.24
3	94.75	100.00	100.00	100.00	100.00
4	87.22	100.00	100.00	100.00	96.43
5	91.80	100.00	100.00	100.00	70.24
6	95.48	100.00	100.00	100.00	91.67
7	93.12	100.00	100.00	75.00	44.05
8	89.33	100.00	100.00	100.00	95.24
9	91.59	100.00	100.00	100.00	100.00
10	89.92	100.00	100.00	100.00	94.05
11	99.26	100.00	100.00	100.00	100.00
12	90.00	100.00	100.00	100.00	96.43
13	97.93	100.00	100.00	100.00	96.43
14	91.60	100.00	100.00	100.00	100.00
15	80.50	93.33	100.00	100.00	57.14
16	87.75	100.00	100.00	100.00	100.00
17	94.44	100.00	100.00	100.00	95.24
18	83.89	100.00	100.00	100.00	92.86
19	97.78	100.00	100.00	100.00	82.14
20	85.61	100.00	100.00	100.00	83.33
21	95.71	100.00	100.00	100.00	92.86
22	94.39	100.00	100.00	100.00	96.43
23	81.91	100.00	100.00	100.00	95.24
24	80.50	100.00	100.00	100.00	95.24
25	90.25	100.00	100.00	100.00	66.67
26	96.67	100.00	100.00	100.00	95.24
27	88.04	100.00	100.00	100.00	98.81
28	91.18	100.00	100.00	100.00	95.24
29	100.00	100.00	100.00	100.00	88.10
30	–	100.00	100.00	100.00	57.14
31	–	100.00	100.00	100.00	91.67
32	–	100.00	100.00	100.00	95.24
33	–	100.00	100.00	100.00	100.00
34	–	100.00	100.00	100.00	72.62
35	–	100.00	100.00	100.00	85.71
36	–	100.00	100.00	100.00	92.86
37	–	100.00	100.00	100.00	94.05
38	–	100.00	66.67	100.00	100.00
39	–	100.00	100.00	100.00	95.24