AgentEconomist: An End-to-End Agentic System for Translating Economic Intuitions into Executable Computational Experiments

Anonymous ACL submission

Abstract

A long-standing challenge in economics lies not in the lack of intuition, but in the difficulty of translating intuitive insights into verifiable research. To address this challenge, we introduce **AgentEconomist**, an end-to-end interactive system designed to translate abstract intuitions into executable computational experiments. Grounded in a domain-specific knowledge base covering over 8,700 high-quality academic papers, the system employs a multi-agent architecture. Specifically, an Idea Development Agent and an Experimental Design Agent collaborate to formulate theoretically grounded hypotheses and draft experimental protocols. An Experimental Execution Agent then conducts the designed experiments, forming a closed-loop workflow that supports the rapid development, validation, and iterative refinement of economic intuitions. Through extensive experiments involving human expert evaluation and large language models (LLMs) as judges, we show that the system generates research ideas with stronger literature grounding and higher novelty and insight than state-of-the-art generic LLMs. Overall, AgentEconomist adopts a human-AI collaboration paradigm that enables researchers to focus on high-level intuitions, while delegating the labor-intensive processes of translation and computational execution to agents[1].

1 Introduction

Scientific research often begins with a spark of intuition, a pre-formal, qualitative understanding of how underlying mechanisms in the world function, which serves as the starting point of any formal ideas and hypotheses (Polanyi, 2009; Popper, 2005). In economics, such intuitions typically concern how agents respond to incentives (Becker, 1976), how institutions shape behavior (North, 1990), and how these interactions give rise to macro-level outcomes (Schelling, 2006). To move beyond mere conjecture, economists traditionally rely on theoretical derivation (Samuelson, 1948) or field experiments (Harrison and List, 2004; Duflo and Banerjee, 2011). However, as illustrated in Figure 1, this traditional workflow is inherently slow and resource-intensive, thereby limiting the scope of ideas a researcher can explore. Computational experiments, based on agent-based modeling (ABM), are a well-established methodology in economics and finance for simulation-based analysis (Tesfatsion, 2006; LeBaron, 2006; Farmer and Foley, 2009) and are frequently employed in teaching to convey complex economic mechanisms (Tisue et al., 2004; Epstein and Axtell, 1996; Railsback and Grimm, 2019). Yet, while such simulation environments provide a powerful medium for verifying economic insights, their inherent complexity often hinders their practical feasibility in research workflows (Axtell and Farmer, 2025).

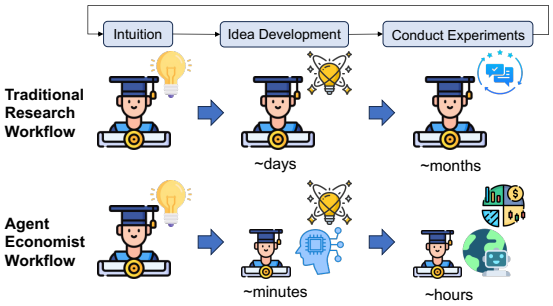

Figure 1: Workflow Comparison. Traditional research (top) suffers from long iteration cycles. In contrast, AgentEconomist (bottom) bridges intuition and execution, enabling rapid hypothesis verification via automated simulation.

This barrier between intuition and computational experiments manifests in three distinct dimensions in traditional modeling workflows. First, the process of idea development is often implicit. Novices struggle to systematize their intuition into con-

[1] https://anonymous.4open.science/r/AgentEconomist-CB75/

crete modeling choices because key methodological knowledge—such as selecting plausible mechanisms or assumptions—remains "tacit knowledge" locked in the minds of experts, rarely explicit in textbooks. Second, turning an idea into a computational experiment is itself challenging. Even when a researcher possesses a clear idea, translating it into an executable simulation experiments involves navigating complex codebases, creating a friction that slows down hypothesis testing. Third, and perhaps most critically, the research process lacks systematic experience accumulation. Scientific discovery is an iterative path of refinement, yet the rationale behind specific iterations, why a parameter was tweaked or a hypothesis rejected, is poorly organized alongside the simulation experiments. This loss of *epistemic context* forces researchers to rely on fragmented memory rather than a structured history of inquiry, making it difficult to learn from past failures or incrementally refine ideas.

Recent advances in leveraging LLM agents for autonomous knowledge discovery have attempted to streamline scientific workflows, yet they reveal substantial limitations when assisting researchers in such complex domains. First, existing efforts often target isolated stages, e.g., optimizing either hypothesis generation (Gottweis et al., 2025) or code execution (Swanson et al., 2025), rather than supporting the seamless workflow required to link literature to simulation. Second, holistic frameworks like *The AI Scientist* (Lu et al., 2024) tend to prioritize "outcome automation" (e.g., generating a final manuscript) over "process support". By treating research as a black box to be automated, they neglect the critical need for *sense-making*: helping researchers structure their thinking and understand the "why" behind modeling decisions. Finally, current systems largely pursue generic applicability at the expense of domain depth, lacking the specific grounding required for economics, where research is deeply rooted in theoretical frameworks and institutional constraints.

To address these challenges, we focus on economics as a representative domain and introduce **AgentEconomist**, an end-to-end interactive research copilot for economic simulation. Built on top of AgentEconomy, a comprehensive agent-based economic simulator, AgentEconomist is designed to support the entire intuition-to-experiment workflow rather than replacing the researcher. Crucially, the system grounds idea development in a large-scale corpus of economic literature, leveraging over 8,700 academic papers from top-tier journals to make tacit theoretical knowledge explicit and accessible. The workflow is decomposed into three specialized roles. An *Idea Development Agent* supports literature-grounded sense-making by retrieving relevant economic studies and synthesizing mechanisms, assumptions, and variables. An *Experimental Design Agent* formalizes these ideas into testable hypotheses and executable experiment specifications. An *Experimental Execution Agent* operationalizes the design by running simulations and returning structured results. Additionally, A structured memory module preserves theoretical context, experimental decisions, and outcomes across iterations, enabling cumulative and context-aware reasoning.

Evaluating such open-ended research assistants presents a unique challenge, as standard benchmarks cannot capture the nuance of scientific reasoning. Therefore, we design a rigorous **mixed-methods evaluation protocol**. First, to assess the quality of Idea Generation, we employ a dual-evaluation framework where both advanced LLMs and human experts score generated hypotheses across eight distinct dimensions (e.g., economic soundness, novelty, and feasibility). Second, to validate the system's utility in real-world workflows, we conduct a holistic user study. Through questionnaires and semi-structured interviews, we gather qualitative feedback on the researchers' experience, focusing on how the system affects their cognitive load, sense of agency, and overall research efficiency. Across both evaluation protocols, our results show that AgentEconomist consistently outperforms strong baselines on the core dimensions of hypothesis novelty and literature grounding, while also producing hypotheses that are more readily operationalizable in simulation-based settings. The primary contributions of this work include:

- **A human-AI collaboration workflow for bridging the intuition-execution gap.** We conceptualize the translation of economic intuition into computational experiments as a collaborative process. By explicitly modeling the transition from implicit sense-making to operational execution, we provide a structured workflow that lowers the technical barriers for economic modeling.
- **An end-to-end system architecture for grounded experimentation.** We introduce *AgentEconomist*, a unified system that integrates

a retrieval-augmented knowledge base (8,700+ papers) with specialized agents. This design ensures that generated hypotheses are theoretically sound and that designed experiments are executable, forming a closed loop for iterative discovery.

- **Empirical validation via mixed-methods assessment.** We conduct quantitative evaluations across 8 quality dimensions and qualitative user interviews, demonstrating that AgentEconomist effectively produces higher-quality research ideas and significantly streamlines the research workflow, empowering users to explore complex scenarios previously inaccessible to novices.

2 Related Works

2.1 Task-Specific Scientific Assistants

Recent research has focused on augmenting specific stages of the research lifecycle. In the realm of data management and deep research (Huang et al., 2025), specific systems like *ChatPD* (Xu et al., 2025) and *SciSciGPT* (Shao et al., 2025b) have emerged to automate dataset discovery and literature analysis, though benchmarks like *DATASETRESEARCH* (Li et al., 2025) reveal that current agents still struggle significantly with out-of-distribution dataset demands. For ideation, methods have evolved from iterative prompting (Gottweis et al., 2025) to multi-agent frameworks like *VIRSCI* (Su et al., 2025), which simulate team collaboration to enhance idea novelty. In the evaluation phase, approaches vary by technical architecture: *ReviewRL* (Zeng et al., 2025b) employs reinforcement learning to optimize feedback quality, while *ReviewAgents* (Gao et al., 2025b) and *DeepReview* (Zhu et al., 2025) utilize multi-agent collaboration and chain-of-thought reasoning to bridge the gap with human peer reviews. Crucially, however, assessments like *IdeaBench* (Guo et al., 2025) expose a paradox: while LLMs generate highly novel ideas, they often lack practical feasibility. This highlights a limitation of such "open-loop" tools: they produce concepts without the agency to validate them. In contrast, AgentEconomist bridges this gap by coupling literature-grounded ideation with an execution toolbox, ensuring abstract intuitions are operationalized into testable simulations.

2.2 Autonomous Research Systems

A parallel stream of work aims to replicate the full role of a human researcher through end-to-end automation (Hu et al., 2025). Building on the code-generation capabilities of *The AI Scientist* (Lu et al., 2024), recent systems have achieved significant breakthroughs: *CycleResearcher* (Weng et al., 2024) introduced a closed-loop "research-review-revise" mechanism, while *DeepScientist* (Weng et al., 2025) incorporated Bayesian optimization for long-term discovery. To support these autonomous agents, infrastructure like *ToolUniverse* (Gao et al., 2025a) has emerged to standardize tool usage. Furthermore, recognizing the complexity of scientific inquiry, frameworks like *OmniScientist* (Shao et al., 2025a) and *MirrorMind* (Zeng et al., 2025a) have expanded this vision by integrating diverse agents to model professional research workflows and collaborative reasoning. However, critical gaps remain when applying these generalist systems to economic inquiry. Existing models often function as black boxes that prioritize outcome automation over process support, denying novices the opportunity for sense-making. Moreover, they lack the deep domain grounding required to map theoretical constructs to high-dimensional simulation parameters. To address this, AgentEconomist positions itself as an interactive co-pilot grounded in a specialized knowledge base of over 8,700 academic papers. By coupling domain depth with rigorous execution, it ensures that scientific discovery remains a transparent, human-aligned process.

3 Method

3.1 Design Rationale: Human–Agent Complementarity

Economic research involves heterogeneous cognitive demands. Human researchers excel at forming high-level intuitions, exercising normative judgment, and deciding when an explanation or result is sufficient. In contrast, transforming these intuitions into executable experiments requires systematic literature grounding, formal specification, and reliable execution—tasks that are labor-intensive and error-prone when performed manually.

AgentEconomist is designed around this complementarity. The system assigns abstraction, judgment, and goal-setting to the human researcher, while delegating literature-grounded sense-making, experiment formalization, and execution management to automated components. This separation

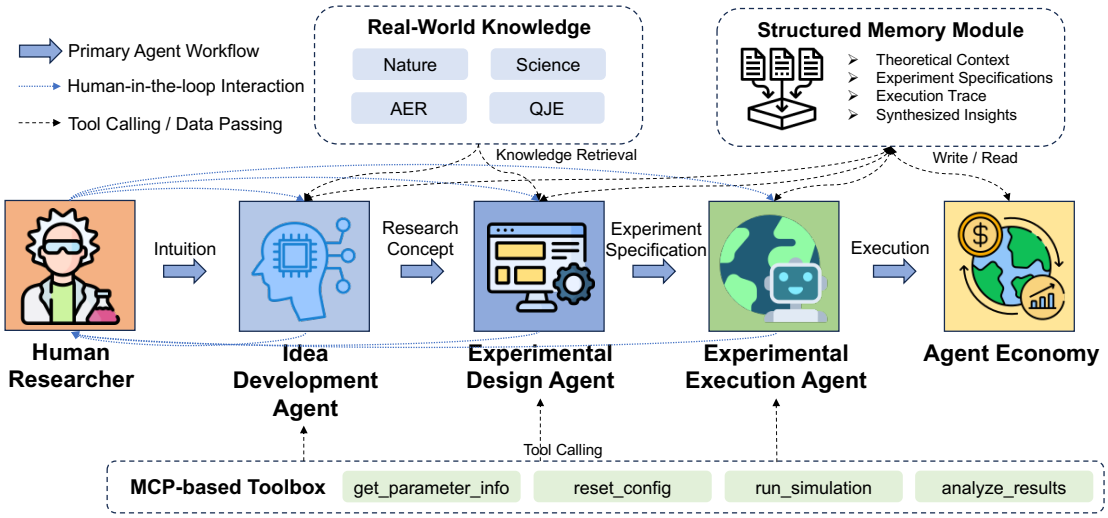

Figure 2: Overview of the AgentEconomist framework. The system decomposes the intuition-to-experiment workflow into literature-grounded ideation, experimental formalization, and execution. These stages are implemented through specialized agents that coordinate via a structured memory module and access the simulator through an MCP-based toolbox, with the human researcher in the loop.

allows each part of the workflow to be handled by the entity best suited for it, while preserving human control over research direction and interpretation.

3.2 The AgentEconomist Framework

Building on the design rationale above, we introduce the AgentEconomist framework shown in Figure 2. The framework operationalizes the intuition-to-experiment workflow through a modular agent architecture, supported by a shared infrastructure layer.

3.2.1 Agent Core

The cognitive workflow is decomposed into three specialized agents corresponding to key stages of economic research:

- **Idea Development Agent.** Responsible for literature-grounded sense-making. It translates human intuition into structured research concepts and candidate hypotheses by retrieving relevant economic literature via retrieval-augmented generation (RAG) and synthesizing key mechanisms and variables (Fig. 3b–c). Beyond initial ideation, this agent also supports iterative verification by checking whether emerging simulation results are consistent with the proposed hypotheses and by guiding hypothesis refinement across iterations (Fig. 3f).

- **Experimental Design Agent.** Responsible for formalization. Given a research concept, it translates abstract hypotheses into executable experimental plans by specifying concrete simulation parameters, defining the number and structure of experiments (e.g., control–treatment or multi-condition setups), and selecting evaluation metrics that align with the simulation environment (Fig. 3d).

- **Experimental Execution Agent.** Responsible for operationalization. It executes experiments in AgentEconomy via the MCP-based toolbox, manages execution states, and returns structured results for downstream analysis and interpretation (Fig. 3e).

3.2.2 Agent Infrastructure

The agent core is supported by three foundational infrastructure components that ensure grounding, continuity, and reliable execution:

- **Real-World Knowledge Base.** To support literature-grounded reasoning, we construct a vector database containing over 8,700 academic papers from top-tier economics and interdisciplinary journals. Retrieved literature supports both idea development and experimental design by making relevant mechanisms, assumptions, and empirical precedents explicit.

- **Structured Memory Module.** The system maintains a persistent memory that records theoretical context, experiment specifications, execution traces, and synthesized outcomes. This enables

context-preserving iteration and cumulative reasoning across multiple research cycles.

- **MCP-based Toolbox.** AgentEconomist interacts with the simulator through a standardized toolbox following the Model Context Protocol (MCP). The toolbox abstracts low-level simulator APIs into semantic actions for parameter inspection, configuration, execution, and analysis, ensuring reliable and reproducible operation.

3.3 Simulation Substrate: AgentEconomy

To support the execution and evaluation of experiments specified by AgentEconomist, we develop AgentEconomy as a computational laboratory for rapid hypothesis verification. Unlike toy models that isolate specific sectors, AgentEconomy provides a comprehensive and flexible agent-based environment designed to capture the ripple effects of economic interventions across interacting components.

Comprehensive Economic Ecosystem. The environment models a closed-loop economic system comprising four core entity types: households, firms, a government, and a bank. These entities interact through two explicitly modeled markets grounded in real-world data:

- **Labor Market:** Matches household skills to firm job requirements, allowing wages and employment to emerge endogenously from supply–demand dynamics.
- **Product Market:** Facilitates the exchange of goods, where consumption is driven by heterogeneous preferences rather than stylized aggregate functions.

This comprehensive coverage enables AgentEconomist to test hypotheses involving multi-stage transmission channels, such as how policy interventions affect household outcomes through firm-level and labor-market adjustments.

Flexible LLM-Driven Behavior. Crucially, agents in AgentEconomy do not rely on rigid, hard-coded heuristics. Instead, they operate through a hybrid mechanism that combines LLM-based reasoning with empirical grounding. Initialized with microdata (e.g., PSID profiles) to ensure realism, agents reason adaptively about decisions such as labor supply and pricing in response to changing incentives. This flexibility allows the system to simulate plausible responses to novel policy shocks (e.g., AI regulation, universal basic income) that are difficult to encode using fixed behavioral rules.

Testbed for Rapid Experimentation. By combining structural completeness with behavioral flexibility, AgentEconomy functions as a high-fidelity testbed for hypothesis-driven experimentation. It exposes a rich set of tunable parameters (policy levers) and observable metrics (macro-level outcomes), enabling AgentEconomist to translate abstract intuitions into concrete simulation runs and receive immediate feedback on validity and emergent phenomena.

3.4 Human–Agent Interaction

Figure 3 illustrates the interactive workflow of AgentEconomist. The process begins with the human researcher providing a high-level research intuition. The system then supports literature-grounded idea development, experiment formalization, and execution in sequence. Results are presented through visualizations and structured summaries, enabling the researcher to assess outcomes and iteratively refine the inquiry while maintaining full contextual awareness.

4 Experiments

We evaluate AgentEconomist as an end-to-end research copilot for economic simulation, focusing on whether it improves the *intuition-to-experiment workflow*. Our evaluation is guided by three research questions:

- **RQ1:** Can the system translate abstract economic intuitions into well-scoped, testable research ideas?
- **RQ2:** Does the system effectively ground these hypotheses in relevant economic literature and coherent economic mechanisms?
- **RQ3:** Can the system operationalize hypotheses into executable experimental configurations that support iterative, context-preserving exploration?

These questions cannot be adequately evaluated using static datasets or fixed benchmarks. Economic research assistance is inherently interactive, open-ended, and path-dependent: the quality of an experiment depends on how intuitions are refined, grounded, and operationalized over multiple turns, rather than on a single correct output.

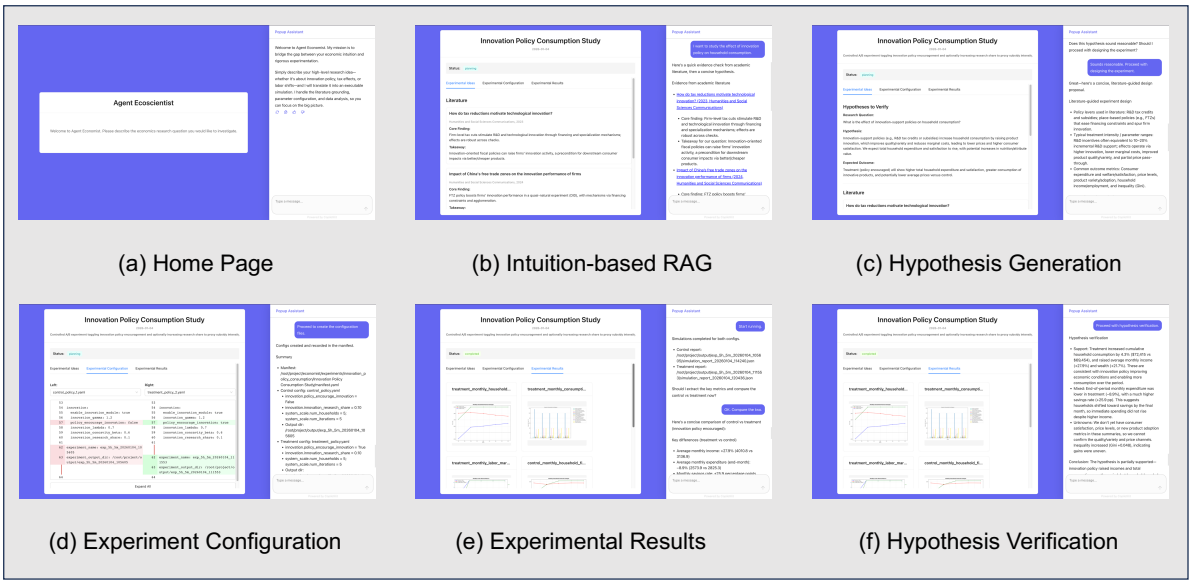

Figure 3: The interactive research workflow of AgentEconomist, illustrating the end-to-end process from intuition input and literature-grounded ideation to experiment configuration, execution, and hypothesis verification.

Accordingly, we adopt a mixed-methods evaluation design that combines controlled comparisons against strong LLM baselines with human-in-the-loop assessments based on real user interactions.

We first describe the experimental setup and data collection procedure (§4.1), then report quantitative results on hypothesis quality across multiple economic dimensions (§4.2), and finally present a qualitative analysis of user feedback to characterize perceived trust, workflow support, and usability limitations (§4.3).

4.1 Experimental Setup

Systems and Participants. We evaluate AgentEconomist against strong general-purpose LLM baselines commonly used as standalone research assistants. AgentEconomist is instantiated with GPT-5 as its backbone model and augmented with domain-specific retrieval, structured memory, and simulator-oriented tooling, as described in §3. Baselines include advanced LLMs such as GPT-5.2 and Gemini 3, used in a vanilla conversational setting without explicit memory or execution support.

We recruited 15 participants with backgrounds in economics, public policy, or related social science fields. Each participant interacted with the system by providing an economic intuition of their own choosing. All participants completed a structured questionnaire; however, due to incomplete submissions, only 14 participants provided full and usable interaction logs for hypothesis-level comparisons.

4.2 Idea Quality and Grounding (RQ1 & RQ2)

This section evaluates whether AgentEconomist improves the quality of generated hypotheses and their grounding in economic literature.

Evaluation Protocol. To isolate hypothesis-level quality from downstream execution variability, we conduct controlled comparisons at the hypothesis generation stage. Both AgentEconomist and baseline LLMs are prompted with identical user intuitions. Generated hypotheses are evaluated along eight dimensions capturing economic rigor, insight, and implementability, using both an LLM-based anonymous referee and human judgments.

Results. Figure 4 summarizes the evaluation results. Across both evaluation protocols, AgentEconomist shows clear advantages on the dimensions most central to the intuition-to-hypothesis translation task. In particular, Literature Grounding and Novelty & Insight exhibit the largest improvements. Under LLM-based judging, Literature Grounding increases from 3.36 to 4.93 and Novelty & Insight from 3.00 to 4.43. Human evaluations show consistent gains in the same dimensions, with Literature Grounding improving from 3.11 to 4.50 and Novelty & Insight from 3.12 to 4.05.

These results are especially significant given our evaluation focus. The primary challenge in economic research assistance is not producing fluent text, but translating vague intuitions into hypothe-

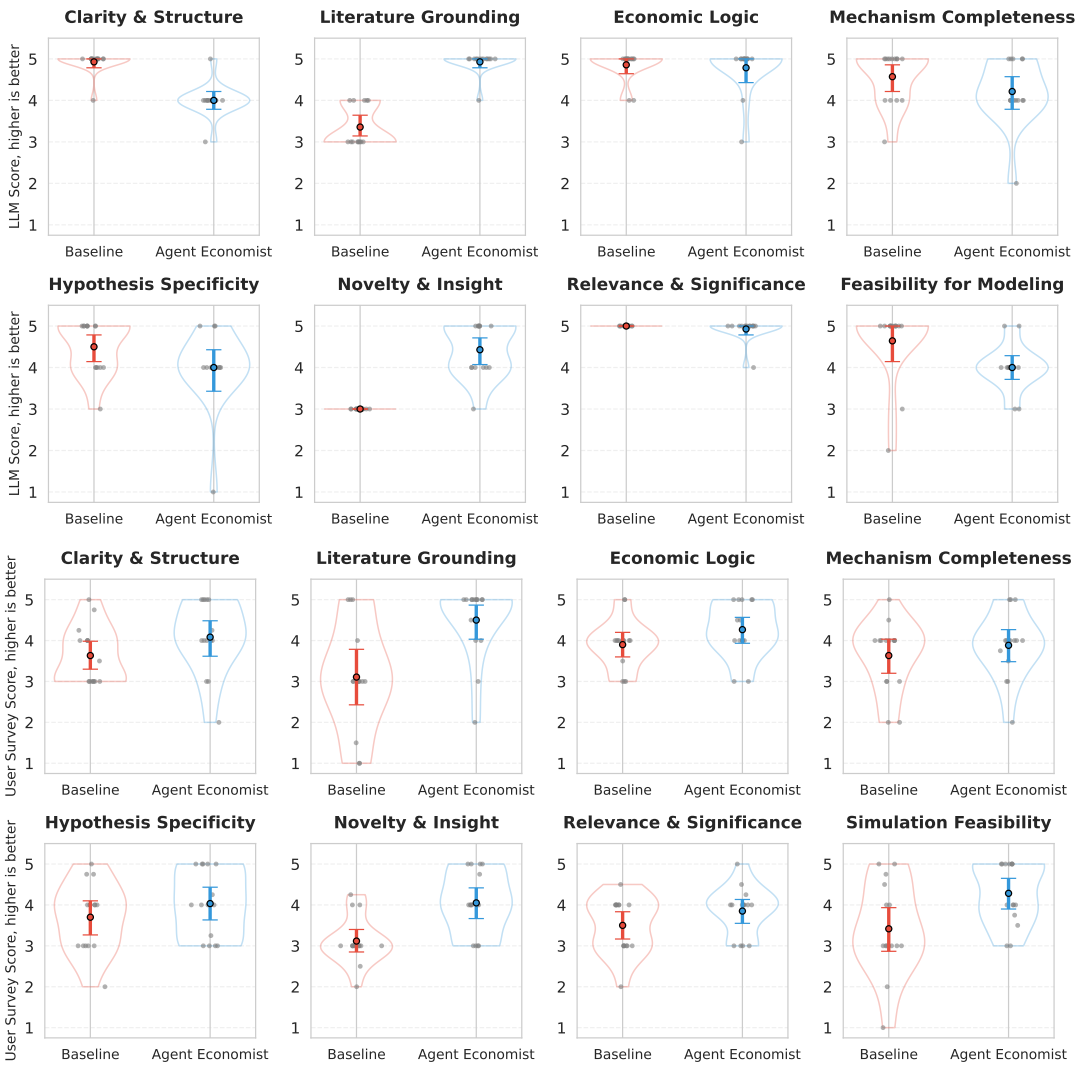

Figure 4: Distribution of hypothesis quality scores across eight evaluation dimensions. **Top**: LLM-based anonymous referee evaluation. **Bottom**: Human evaluation by study participants. Each subplot corresponds to one dimension. Across both judging protocols, AgentEconomist consistently outperforms the baseline, with the largest margins observed in *Literature Grounding* and *Novelty & Insight*.

ses that are both theoretically grounded and intellectually non-trivial. On these core criteria, AgentEconomist achieves substantial and consistent improvements over strong baseline models.

Human–LLM Evaluation Differences. A systematic divergence arises in the Clarity & Structure dimension. While the LLM-based referee slightly favors the baseline, human evaluators rate AgentEconomist comparably or higher. We attribute this difference to distinct evaluation priorities. Baseline outputs tend to be shorter and rhetorically streamlined, whereas AgentEconomist produces denser hypotheses that explicitly articulate mechanisms, assumptions, and literature connections. Although this additional structure may reduce surface-level clarity for automated judges, it aligns more closely with human researchers' needs when developing grounded and novel economic hypotheses.

4.3 Qualitative User Experience Analysis (RQ3)

To examine how users perceive AgentEconomist as an end-to-end research framework, we conduct a formative qualitative analysis of open-ended questionnaire responses. The questionnaire consists of five prompts: Q1 (Perceived Advantages), Q2 (Trust and Credibility), Q3 (Pain Points and Limitations), Q4 (Role Perception), and Q5 (Other Open Feedback).

Not all participants answered every prompt (Q1: $n = 8$, Q2–Q4: $n = 7$, Q5: $n = 3$). Accordingly, our analysis focuses on recurring themes supported by multiple responses rather than esti-

mating population-level prevalence.

Method. We apply an LLM-assisted grounded-theory thematic analysis to aggregate responses and extract emergent themes. Only patterns directly supported by verbatim evidence are reported, and findings are interpreted as experience-level perceptions rather than objective performance claims.

Findings. Participants consistently emphasized grounded trust, attributing increased confidence to literature-backed reasoning enabled by the RAG module. A second dominant theme is operationalization support: users valued the system's ability to translate vague intuitions into simulator-aligned experimental configurations. Third, respondents highlighted mechanistic scaffolding, noting that AgentEconomist more explicitly articulated causal and behavioral chains than generic LLM outputs. Together, these factors contributed to a perceived role shift from a conversational assistant toward a research assistant or collaborative expert spanning hypothesis formulation and experiment setup.

Pain Points and Implications. Feedback also revealed usability bottlenecks, including execution latency, limited process transparency, and occasional instruction-following drift over long interactions. These observations suggest that future work should prioritize improving responsiveness, exposing interpretable intermediate states, and strengthening long-horizon intent preservation.

4.4 Case Study

We illustrate the end-to-end use of AgentEconomist with a representative interaction drawn from a real user session. The user started from a high-level intuition: *whether innovation-support policies increase household consumption*. Given this underspecified query, the system first retrieved and summarized relevant economic literature, including studies on R&D tax incentives, place-based innovation policies, and production–consumption linkages, grounding the inquiry in established evidence.

Conditioned on this context, AgentEconomist formulated a concrete hypothesis that innovation policies raise household consumption through firm innovation, price effects, and income channels. The system then designed a controlled A/B experiment by toggling a policy parameter (`innovation.policy_encourage_innovation`), holding market structure constant, and specifying treatment and control configurations. Key parameters—including the number of households, simulation horizon, and innovation research share—were explicitly exposed and iteratively adjusted by the user.

After executing the simulations, the system automatically extracted and compared outcome metrics. The treatment condition exhibited higher cumulative household consumption (+4.3%), substantially higher income (+27.9%) and wealth (+21.7%), alongside a higher savings rate and slightly increased inequality. AgentEconomist synthesized these results into a structured verification report, concluding that the hypothesis was partially supported and recommending further analysis.

The interaction terminated once the user was satisfied with the explanation and next-step suggestions. This case demonstrates that AgentEconomist supports the full research workflow—from intuition, to literature grounding, hypothesis generation, and experimental design and iteration—within a unified and context-aware framework.

5 Conclusion

We presented AgentEconomist, an end-to-end agentic system that supports economic research by translating abstract intuitions into executable computational experiments. The system adopts a human-in-the-loop design that decomposes the intuition-to-experiment workflow into literature-grounded idea development, experiment formalization, and execution, implemented by specialized agents. Empirical evaluation shows that AgentEconomist improves hypothesis quality on the dimensions most critical to economic inquiry, particularly literature grounding and novelty of insight, while remaining effective in interactive research settings. Throughout this process, the system remains controllable executions and iterations explicitly mediated by the user, mitigating potential risks associated with autonomous behavior. More broadly, this work highlights the importance of epistemic scaffolding in scientific AI systems, where supporting theory-grounded reasoning and iterative sense-making is often more valuable than end-to-end automation. We hope this perspective informs future work on domain-grounded, interactive agents for scientific discovery.

Limitations

This work has several limitations. First, our evaluation focuses on the intuition-to-experiment workflow in an agent-based economic simulation, and does not assess performance on real-world policy deployment or empirical data analysis. Second, while the user study provides evidence of the system's usefulness in interactive research settings, the number of participants is limited and may not capture the full diversity of economic research practices. Third, the effectiveness of AgentEconomist depends on the coverage and quality of the underlying literature corpus and simulation environment; domains or research questions that fall outside these resources may be less well supported. Finally, although the system maintains structured memory across interactions, long-horizon alignment and execution efficiency remain constrained by current LLM capabilities and system latency. Future work may address these limitations by expanding empirical evaluation settings, incorporating larger and more diverse user populations, and improving system robustness and scalability.

References

Robert L Axtell and J Doyne Farmer. 2025. Agent-based modeling in economics and finance: Past, present, and future. *Journal of Economic Literature*, 63(1):197–287.

Gary S Becker. 1976. *The economic approach to human behavior*, volume 803. University of Chicago press.

Esther Duflo and Abhijit Banerjee. 2011. *Poor economics*, volume 619. PublicAffairs New York.

Joshua M Epstein and Robert Axtell. 1996. *Growing artificial societies: social science from the bottom up*. Brookings Institution Press.

J Doyne Farmer and Duncan Foley. 2009. The economy needs agent-based modelling. *Nature*, 460(7256):685–686.

Shanghua Gao, Richard Zhu, Pengwei Sui, Zhenglun Kong, Sufian Aldogom, Yepeng Huang, Ayush Noori, Reza Shamji, Krishna Parvataneni, Theodoros Tsiligkaridis, and 1 others. 2025a. Democratizing ai scientists using tooluniverse. *arXiv preprint arXiv:2509.23426*.

Xian Gao, Jiacheng Ruan, Zongyun Zhang, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. 2025b. Reviewagents: Bridging the gap between human and ai-generated paper reviews. *arXiv preprint arXiv:2503.08506*.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, and 1 others. 2025. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*.

Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Myles Kim, Corey M Williams, Stefan Bekiranov, and Aidong Zhang. 2025. Ideabench: Benchmarking large language models for research idea generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5888–5899.

Glenn W Harrison and John A List. 2004. Field experiments. *Journal of Economic literature*, 42(4):1009–1055.

Ming Hu, Chenglong Ma, Wei Li, Wanghan Xu, Jiamin Wu, Jucheng Hu, Tianbin Li, Guohang Zhuang, Jiaqi Liu, Yingzhou Lu, and 1 others. 2025. A survey of scientific large language models: From data foundations to agent frontiers. *arXiv preprint arXiv:2508.21148*.

Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Huichi Zhou, Meng Fang, Linyi Yang, Xiaoguang Li, Lifeng Shang, Songcen Xu, and 1 others. 2025. Deep research agents: A systematic examination and roadmap. *arXiv preprint arXiv:2506.18096*.

Blake LeBaron. 2006. Agent-based computational finance. *Handbook of computational economics*, 2:1187–1233.

Keyu Li, Mohan Jiang, Dayuan Fu, Yunze Wu, Xiangkun Hu, Dequan Wang, and Pengfei Liu. 2025. Datasetresearch: Benchmarking agent systems for demand-driven dataset discovery. *arXiv preprint arXiv:2508.06960*.

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

Douglass C North. 1990. *Institutions, institutional change and economic performance*. Cambridge university press.

Michael Polanyi. 2009. The tacit dimension. In *Knowledge in organisations*, pages 135–146. Routledge.

Karl Popper. 2005. *The logic of scientific discovery*. Routledge.

Steven F Railsback and Volker Grimm. 2019. *Agent-based and individual-based modeling: a practical introduction*. Princeton university press.

Paul Anthony Samuelson. 1948. Foundations of economic analysis. *Science and Society*, 13(1).

Thomas C Schelling. 2006. *Micromotives and macrobehavior*. WW Norton & Company.

Chenyang Shao, Dehao Huang, Yu Li, Keyu Zhao, Weiquan Lin, Yining Zhang, Qingbin Zeng, Zhiyu Chen, Tianxing Li, Yifei Huang, and 1 others. 2025a. Omniscientist: Toward a co-evolving ecosystem of human and ai scientists. *arXiv preprint arXiv:2511.16931*.

Erzhuo Shao, Yifang Wang, Yifan Qian, Zhenyu Pan, Han Liu, and Dashun Wang. 2025b. Sciscigpt: advancing human–ai collaboration in the science of science. *Nature Computational Science*, pages 1–15.

Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, and 1 others. 2025. Many heads are better than one: Improved scientific idea generation by a llm-based multi-agent system. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28201–28240.

Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. 2025. The virtual lab of ai agents designs new sars-cov-2 nanobodies. *Nature*, 646(8085):716–723.

Leigh Tesfatsion. 2006. Agent-based computational economics: A constructive approach to economic theory. *Handbook of computational economics*, 2:831–880.

Seth Tisue, Uri Wilensky, and 1 others. 2004. Netlogo: A simple environment for modeling complexity. In *International conference on complex systems*, volume 21, pages 16–21. Boston, MA.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2024. Cycleresearcher: Improving automated research via automated review. *arXiv preprint arXiv:2411.00816*.

Yixuan Weng, Minjun Zhu, Qiujie Xie, Qiyao Sun, Zhen Lin, Sifan Liu, and Yue Zhang. 2025. Deepscientist: Advancing frontier-pushing scientific findings progressively. *arXiv preprint arXiv:2509.26603*.

Anjie Xu, Ruiqing Ding, and Leye Wang. 2025. Chatpd: An llm-driven paper-dataset networking system. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5106–5116.

Qingbin Zeng, Bingbing Fan, Zhiyu Chen, Sijian Ren, Zhilun Zhou, Xuhua Zhang, Yuanyi Zhen, Fengli Xu, Yong Li, and Tie-Yan Liu. 2025a. Mirrormind: Empowering omniscientist with the expert perspectives and collective knowledge of human scientists. *arXiv preprint arXiv:2511.16997*.

Sihang Zeng, Kai Tian, Kaiyan Zhang, Yuru Wang, Junqi Gao, Runze Liu, Sa Yang, Jingxuan Li, Xinwei Long, Jiaheng Ma, and 1 others. 2025b. Reviewrl: Towards automated scientific review with rl. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16942–16954.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025. Deepreview: Improving llm-based paper review with human-like deep thinking process. *arXiv preprint arXiv:2503.08569*.

A Experiment Details

A.1 Hypothesis Quality Dimensions

Participants were asked to evaluate each generated hypothesis along the following eight dimensions. For each dimension, participants were instructed to provide an independent score based on their own judgment, focusing on the hypothesis content rather than writing style or presentation.

- **Clarity & Structure:** To what extent is the hypothesis clearly stated, logically structured, and free of ambiguous or vague claims?

- **Literature Grounding:** To what extent is the hypothesis consistent with established economic theory and plausibly grounded in relevant academic literature, without hallucinated or misused references?

- **Economic Logic:** To what extent does the hypothesis align with core economic principles such as incentives, constraints, and rational behavior?

- **Mechanism Completeness:** To what extent does the hypothesis specify a coherent causal or behavioral mechanism linking the intervention to the outcome?

- **Hypothesis Specificity:** To what extent is the hypothesis concrete, testable, and operational, rather than generic (e.g., clearly specifying agents, directions, or conditions)?

- **Novelty & Insight:** To what extent does the hypothesis provide non-trivial or original insights beyond well-known or widely accepted economic claims?

- **Relevance & Significance:** To what extent does the hypothesis address an economically or policy-relevant question with potential real-world importance?

- **Simulation Feasibility:** To what extent can the hypothesis be realistically implemented within an agent-based simulation framework without relying on unobservable or ill-defined variables?

A.2 Scoring Protocol

Each dimension is scored on a 5-point Likert scale: 1 = Very poor, 3 = Acceptable, 5 = Excellent. Judges are instructed to base scores solely on the hypothesis-generation content and to ignore writing style, verbosity, or downstream analysis.

A.3 Qualitative User Experience Analysis Protocol

This qualitative evaluation is designed as a formative assessment to capture participants' subjective experiences when interacting with AgentEconomist as a research framework. Participants are asked to provide free-text responses based on their own interaction trajectories, without assuming complete task coverage or uniform interaction length.

Evaluation Dimensions. Open-ended questions are organized around four aspects of user experience:

- **Perceived Advantages:** Key differences compared to using a general-purpose LLM.
- **Trust and Credibility:** Factors influencing confidence in generated hypotheses and experimental designs.
- **Pain Points and Limitations:** Usability issues, missing features, or interaction difficulties.
- **Role Perception:** How users conceptualize the system (e.g., search engine, research assistant, collaborative expert).

Response Format. Participants provide short free-text answers (approximately 100 words per question). Responses are anonymized and aggregated for analysis.

Analysis Protocol. Aggregated responses are analyzed using an LLM-assisted grounded-theory workflow. The analysis extracts emergent themes through iterative coding and grouping, with final reporting restricted to themes supported by multiple responses and direct textual evidence.

A.4 Participant Background and Compensation

All participants in the user study were doctoral students actively working in the area of economic modeling or simulation-based economic research. Each participant had at least six months of prior research experience related to economic simulation or computational economics, ensuring sufficient domain knowledge to meaningfully evaluate the system.

Participants were compensated for their time at a rate commensurate with their academic background and local standards. We consider the payment to be adequate given the participants' demographic and level of expertise, and the study involved no deceptive practices or sensitive data collection. All participants provided informed consent prior to participation and were explicitly informed about how their interaction data would be used for research and evaluation purposes.

B Prompts Demonstration

In this section, we provide the prompts used in our evaluation pipeline. Each {...} placeholder in the templates will be substituted with the corresponding study materials (e.g., hypothesis excerpts, anonymized IDs, and aggregated free-text responses). For implementation details and exact instantiations, please refer to our code release.

Table 1 presents the LLM-based referee prompt used for hypothesis-quality scoring. Table 2 presents the grounded-theory prompt used for LLM-assisted thematic analysis of open-ended feedback.

C AI Assistance Disclosure

AI-based tools were used in the preparation of this work for code development assistance and for grammatical and stylistic editing of the manuscript. All scientific content, experimental design, results, and conclusions were conceived, implemented, and verified by the authors.

Role: Anonymous Economics Referee (LLM-as-a-Judge)
You are acting as an anonymous, neutral, and rigorous economics referee.
Your task is to evaluate and compare the quality of **hypothesis-generation content** produced by two systems in response to the same economic question.

[Critical Scope Restriction]
Evaluate **ONLY the generated hypotheses themselves**. Do NOT consider:
- downstream analysis, simulations, policy discussion, extensions, or follow-up content.

[Evaluation Materials]
- Materials are provided as images or PDF pages.
- First {N} pages belong to System A; following {M} pages belong to System B.
- Output order does NOT imply priority. Ignore system identity.

[Evaluation Principles]
1) Content-only evaluation: ignore style, verbosity, tone, or rhetorical confidence.
2) Order invariance: do not favor earlier or later materials.
3) Model blindness: do not infer system identity or sophistication.
4) Referee-level standards: judge as if reviewing an economics paper's hypotheses.

[Dimensions: score each 1–5 (integer)]
(1) Clarity & Structure
(2) Literature Grounding & Factual Plausibility
(3) Economic Logic / Soundness
(4) Mechanism Completeness (at hypothesis level)
(5) Hypothesis Specificity
(6) Novelty & Insight
(7) Relevance & Significance
(8) Feasibility for Modeling or Simulation

[Required Output Format: Markdown Only]
- A dimension-wise score table comparing System A vs System B with brief justification.
- Overall assessment (200 words).
- Bias & scope compliance check:
a) evaluated hypotheses only? (Yes/No)
b) ignored identity and order? (Yes/No)
c) any dimensions hard to score due to insufficient detail?

Table 1: Prompt template for LLM-based hypothesis-quality judging (anonymous economics referee). Each {...} placeholder is instantiated with the corresponding evaluation materials.

Role: Social Science Researcher (Grounded Theory)
You are a social science researcher familiar with qualitative methods.
Please use grounded theory (Grounded Theory) to conduct a systematic thematic analysis on the following interview materials.

[Methodological Constraint]
Remain highly sensitive to the data and avoid pre-set theoretical assumptions.
Ensure conclusions are grounded in the interview materials themselves.

[Required Output Structure]
- Core themes
- Theme summary
- Verbatim example sentences

[Interview Materials]
{Paste aggregated anonymized responses here, e.g., P1:, P2:, ...}

Table 2: Prompt template for LLM-assisted grounded-theory thematic analysis of participants' open-ended feedback.