Position: Assumption Search, Not Proof Search: AI Assistance for Theory-Oriented Social Science

Anonymous Author(s) [1]

Abstract

This position paper argues that AI-for-math paradigms do not transfer directly to theory-oriented social science. Although theoretical social science—especially areas such as theoretical economics—often contains substantial mathematical derivation and proof, its central research bottleneck is frequently not proof search under strong verification. Instead, progress often depends on assumption search: identifying assumptions, mechanisms, boundary conditions, and interpretations that are coherent, meaningful, and disciplinarily acceptable under weak external validation. This difference has important implications for AI system design. We argue that AI assistance in such settings should be organized around human-in-the-loop workflows that support problem framing, assumption proposal, verification, branching, and scholarly adjudication, rather than around proof-centric automation alone. We use Adrasteia as a concrete instantiation of this position and present limited empirical evidence as supporting illustration rather than definitive validation. Our goal is not to claim that theory-oriented social science can be reduced to workflow engineering, but to argue that future AI systems for these domains should treat human judgment and assumption search as first-class design principles. The paper is centered on economics and offers only cautious extrapolation to adjacent theory-oriented social-science settings.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author(s) <anonymous@example.com>.

The 3rd *AI for Math Workshop* at the 43rd International Conference on Machine Learning (ICML), Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

1. Introduction

1.1. Why Existing General LLMs and AI-for-Math Paradigms Do Not Directly Fit Theory-Oriented Research in Social Science

Recent progress in AI for reasoning and research has been substantial. General-purpose language models are increasingly capable at writing, summarization, coding assistance, and broad question answering, while research-oriented systems such as *Aletheia*, *AI co-scientist*, and *Aristotle* organize reasoning as iterative workflows rather than one-shot prompting (1–3). These advances matter for theory-oriented social science, but they do not by themselves determine how AI should be designed for that setting.

Position. AI-for-math paradigms do not transfer directly to theory-oriented social science. Although theory-oriented social science—especially areas such as theoretical economics—often contains substantial mathematical derivation and proof, its central research bottleneck is frequently not proof search under strong verification. Instead, the harder step is often *assumption search*: identifying assumptions, mechanisms, boundary conditions, and interpretations that are coherent, meaningful, and disciplinarily acceptable under weak external validation.

This distinction matters precisely because recent AI progress in mathematics has been so strong. There, frontier systems have flourished in verifier-rich environments in which progress can be anchored to external standards of correctness. *AlphaProof*, for example, is explicitly built around the verification power of formal systems: its Nature paper argues that the inherent verification capabilities of formal systems provide a foundation for trusted AI reasoning, and the system searches for proofs inside the *Lean* theorem prover environment (4). *Lean* itself is officially described as an interactive theorem proving environment for developing and verifying proofs (6). In parallel, *AlphaGeometry* demonstrated that structured reasoning over geometry problems can reach Olympiad-level performance (5), while *FunSearch* showed that when a language model is paired with a systematic evaluator, search over candidate programs can yield new mathematical and algorithmic discoveries (7). The common feature across these successes is not merely

stronger language generation, but the presence of strong evaluators, formal checks, or tightly specified success criteria.

Crucially, these systems are no longer confined to competition mathematics. Public descriptions of *Aletheia* emphasize that the transition from Olympiad-level problem solving to professional mathematical research requires navigating vast literatures and constructing long-horizon arguments (1). The reported milestones already include fully autonomous generation of a research paper on eigenweights in arithmetic geometry, human–AI collaboration on bounds for independent sets of interacting particles, and a semi-autonomous large-scale evaluation of 700 open problems in Bloom's Erdős conjecture database, including autonomous solutions to four open questions (1; 8; 9). More broadly, recent case studies with advanced Gemini-based systems document the use of frontier models not only to solve open problems but also to refute conjectures and generate new proofs across theoretical computer science, economics, optimization, and physics (8). In short, proof-centric paradigms can scale impressively when the task itself admits strong external checks.

Theory-oriented social science differs structurally on exactly this dimension, even if a small number of adjacent theory-oriented settings share parts of the same problem structure. Mathematics typically begins with a fixed proposition, conjecture, or theorem statement and searches over proof techniques. By contrast, theoretical model building in economics and closely related traditions often begins with a target phenomenon, mechanism intuition, explanatory narrative, or desired claim, and then searches over assumptions under which that account becomes coherent and persuasive. The bottleneck is therefore often not deriving conclusions from assumptions, but identifying assumptions that are theoretically meaningful, empirically defensible where relevant, and disciplinarily acceptable. In mathematics, formal proof assistants can serve as a universal external criterion of success (4; 6). In economics-centered theory work, there is no corresponding checker for whether assumptions are natural, interesting, publishable, or intellectually convincing to other researchers. The central challenge is thus not simply proof search, but what we may call *assumption search* under much weaker external validation.

The implication is not that verification is unimportant, nor that theory-oriented social science can be reduced to workflow engineering. It is that AI assistance in these domains should be organized around human-in-the-loop workflows that support problem framing, assumption proposal, verification, branching, and scholarly adjudication, rather than around proof-centric automation alone. *Adrasteia* should therefore be read as one concrete instantiation of this broader claim, not as the claim itself. More generally, the field should treat human judgment and assumption search as first-class design principles when building AI systems for theory-oriented social science.

1.2. Supporting Claims from Current LLM Use in Economics and Adjacent Social-Science Research

We do not review this literature for its own sake. Instead, we extract a small number of claims from recent work in economics and adjacent social-science settings that support the position advanced above (16).

Claim 1: The literature has already moved well beyond writing assistance. The earliest wave of adoption framed language models as tools for *cognitive automation in research*: summarization, explanation, translation, rewriting, drafting, literature organization, and basic coding assistance (10). LLMs were then rapidly embedded into empirical and econometric workflows as tools for measurement, annotation, extraction, and classification, with validity tied to explicit research-design safeguards rather than to autonomous economic judgment (13). More recent work pushes further toward reasoning support, theoretical derivation, research design, critical feedback, model revision, and agentic research assistance, with Korinek's later updates explicitly distinguishing *reasoning models* and *agentic systems* from earlier chatbot use (11; 12). In other words, the relevant question is no longer whether LLMs can help economists write, but where their limits appear as tasks move closer to frontier research judgment.[1]

Claim 2: LLM use in economics has bifurcated between research assistance and artificial agents. A parallel literature studies LLMs not as tools for the researcher, but as economic and social agents in their own right, from *Homo Silicus* to LLM-based survey and simulation frameworks (14; 18). This matters here mainly as scope: once LLMs enter economics in both roles, the relevant question for the present paper is not whether they belong in the field, but how they should be organized when the task is theory-oriented research rather than agent simulation.

Claim 3: As one approaches the research frontier, the binding constraint becomes discipline-specific judgment. The most important lesson from the literature is not simply that LLM use is growing, but that the remaining bottleneck has shifted. Current systems may already be strong at writing, coding, literature synthesis, and increasingly at deep

[1]Recent working papers also suggest that frontier LLMs are already being used for substantive modeling tasks, including theoretical model formulation, comparative-statics derivation, and debugging of quantitative implementations. A concrete example is *Trade Policy Ambiguity* (NBER Working Paper No. 34408), whose author discloses direct use of ChatGPT 5 Pro in theoretical and computational components of the project (19).

research and workflow automation, but they remain unreliable when pushed to operate at the level of a disciplinary researcher. Korinek repeatedly emphasizes that current systems may misapply theoretical frameworks, reproduce common misconceptions from their training data, or produce superficially plausible but weakly grounded analysis when operating at researcher level (12). Complementing this concern, *EconNLI* provides direct evidence that LLMs are still not sophisticated in economics reasoning and may generate wrong or hallucinated answers in tasks requiring domain-specific causal and inferential understanding (17). Even the newer workflow-oriented literature on economic agents retains strategic human oversight and HITL checkpoints rather than claiming full disciplinary autonomy (15). For the present paper, the implication is straightforward: the frontier problem in theory-oriented social science is not merely better language generation, but better support for judgment over assumptions, identification, interpretation, and disciplinary acceptability.

Taken together, these claims reinforce the broader position of this paper: current progress is real and relevant, but the crucial design problem in theory-oriented social science is how to assist assumption search under weak external validation rather than how to automate proof-like reasoning alone.

Roadmap. The remainder of the paper proceeds as follows. Section 2 clarifies what this position commits us to, what it does not claim, and which design commitments follow from it, before locating ADRASTEIA as one concrete instantiation. Section 3 then gives the core argument for why assumption search, rather than proof search, is the right abstraction. Section 4 follows by operationalizing that argument through an illustrative workflow organization for theory-oriented social science. Section 5 then discusses alternative views and scope conditions. Section 6 summarizes the illustrative empirical support, and Section 7 concludes.

2. What This Position Implies for AI-Assisted Theory-Oriented Social Science

2.1. What This Position Is

Our position is that AI assistance for theory-oriented social science should be organized around human-in-the-loop assumption-search workflows rather than treated as a direct extension of proof-centric reasoning systems. In this view, the relevant object is not a single final answer, but an inspectable research process involving problem framing, assumption proposal, critique, revision, and scholarly adjudication.

This position is motivated by the broader shift from one-shot prompting toward research-oriented workflows exemplified by systems such as *Aletheia*, *AI co-scientist*, and *Aristotle* (1–3). But the claim of the present paper is that theory-oriented social science is a different setting: not proof search under a strong checker, but assumption search under human judgment. The commitment is therefore to a workflow-level view of assistance rather than to any single model or backend.

2.2. What This Position Is Not

This position does not claim that theory-oriented social science should become fully autonomous, that proof-centric systems are irrelevant, or that theory-oriented social science can be reduced to workflow engineering. It does not argue for an automatic paper-writing machine, nor does it assume that theory-oriented research can be treated as proving a fixed statement under a universal checker.

It also does not claim that ADRASTEIA is the only correct design. The point is narrower: when assumptions, interpretations, and boundary conditions are centrally at issue, human intervention is not something to be engineered away, because question selection, assumption choice, mechanism interpretation, and conclusion framing are part of the substance of scholarship itself.

2.3. Design Commitments

If this position is right, several design commitments follow for AI-assisted theory-oriented social science.

First, such systems should be **semi-automated**. Models may carry out high-frequency generation, checking, and revision tasks, but they should not govern the process end to end. The aim is to amplify research capacity, not to eliminate the role of the researcher.

Second, **human judgment should be treated as a first-class design constraint rather than fallback supervision**. Human researchers should intervene at key points in the workflow, including problem definition, branch selection, interpretation of verifier outputs, and conclusion adoption. Human intervention is therefore not a fallback for system failure; it is an explicit architectural component of the research process.

Third, **verification should be diagnostic and comparative rather than purely accept/reject**. In these settings, checking should help identify hidden assumptions, weak mechanisms, claim overreach, and alternative branches worth retaining, rather than merely certify success or failure.

Fourth, **outputs should be research objects rather than final answers**. The goal is to produce inspectable, revisable artifacts in which assumptions, alternatives, limitations, and judgment calls remain visible, rather than to terminate the research process with a single fluent conclusion.

Table 1. Design consequences of replacing proof search with assumption search.

Dimension	AI for Math	Assumption-Search View
Starting point	Fixed proposition, conjecture, or theorem statement	Research question, phenomenon, mechanism intuition, or target claim
Search space	Proof techniques and derivation paths	Assumptions, mechanisms, boundary conditions, and interpretive branches
Verifier	Formal checker or tightly specified evaluator	Literature, proof-text, and reviewer-style verification
Main bottleneck	Finding a valid proof	Finding acceptable and meaningful assumptions
Success criterion	Valid proof accepted by the verifier	Acceptable, inspectable research object
Human role	Often secondary after formalization	Central in judgment, branch selection, and conclusion framing

2.4. Adrasteia as a Concrete Instantiation

We use ADRASTEIA as one concrete instantiation of these commitments for theory-oriented research in economics, with only cautious extension to adjacent theory-oriented settings. It is not the claim of the paper itself. Rather, it is an illustrative workflow through which a later section shows what follows at the level of system organization if AI assistance is built around assumption search under human judgment.

3. Why Theory-Oriented Social Science Requires Assumption Search

The central claim of this paper is not merely that mathematics and theory-oriented social science are different domains. It is that proof search is the wrong default abstraction for much of theory-oriented social science. Once the search object, verifier semantics, and success criterion are specified correctly, the design and evaluation targets for AI systems change as well.

3.1. Search Object

Table 1 summarizes the design consequences of replacing proof search with assumption search.

In proof-oriented systems, the search object is typically a derivation path. A proposition is given, and the system searches over proof techniques, intermediate lemmas, and formal manipulations that may connect assumptions to the desired conclusion. This is the setting in which systems such as *Aletheia*, *AlphaProof*, and related theorem-proving workflows operate most naturally (1; 4). Once the proposition is fixed, the main question is how to find a valid route to it.

By contrast, theory-oriented social science often searches over candidate model worlds: assumption packages, mechanism structures, scope conditions, and alternative interpretations. The relevant object is therefore not a proof, but a family of possible explanatory constructions under which a target claim becomes coherent and worth entertaining. This is why systems in this setting must explicitly support problem framing, assumption proposal, and branch management. If the search space consists of candidate model worlds rather than candidate proofs, then the task cannot be reduced to a single forward derivation.

This shift has an important implication. Multiple branches may remain viable for long stretches of the research process. A branch need not be discarded simply because it is not uniquely privileged; it may be retained because it offers a different but still potentially meaningful explanation under a different assumption package. Search is therefore comparative rather than purely eliminative.

Therefore, systems and benchmarks for this setting should be built to surface candidate assumption packages, preserve viable branches, and compare model worlds, rather than score only proof completion or single-answer accuracy.

3.2. Verifier Semantics

The change in search object also changes the role of verification. In proof-oriented settings, the verifier is fundamentally accept/reject in character. Its role is to determine whether a derivation satisfies the formal or tightly specified criterion that defines success. Even when search is heuristic, the semantics of verification remain relatively crisp: the proof is either valid or it is not.

In the setting considered here, verification is not primarily accept/reject. It is diagnostic and comparative. The point is not simply to determine whether a candidate branch has "passed." It is to identify where assumptions are too strong, where mechanism claims are under-motivated, where conclusions overreach the stated setup, where textual proofs contain hidden jumps, and where alternative branches may be more natural or more defensible. Verification therefore guides revision, branching, and human adjudication rather than merely certifying correctness.

This distinction is reflected directly in the workflow stance

defended here. Literature and factual verification assess grounding, representation, and external support. Proof- and text-level checking assess internal consistency over natural-language derivations. Reviewer-style critique adds another layer, asking whether the resulting draft can survive disciplinary scrutiny. The semantics of verification are therefore cumulative and interpretive: they generate structured feedback on the quality of a research object rather than issuing a final binary judgment.

Therefore, evaluation should ask whether verification improves diagnosis, comparison, and revision, not merely whether a checker returns pass/fail outcomes on isolated reasoning steps.

3.3. Success Criterion

The final consequence concerns success itself. In proof-oriented systems, success is naturally defined as finding a valid proof accepted by the relevant checker or evaluator. The output semantics are therefore relatively straightforward: once the proof is accepted, the task is complete.

In the setting considered here, success cannot be defined in this way. The relevant output is not a proof, but an acceptable, inspectable research object. Such an object must include at least four features. First, it must provide an explicit assumption-to-conclusion chain. Second, it must preserve viable alternative branches where these remain substantively meaningful. Third, it must document weaknesses, limitations, or unresolved tensions rather than hiding them behind rhetorical fluency. Fourth, it must preserve a record of human judgment in selection, interpretation, and conclusion framing.

Accordingly, the endpoint is not a single privileged derivation, but a research object that can be inspected, criticized, extended, and possibly subjected to further internal or external review. Success therefore means producing something that is not only internally coherent, but also interpretable, revisable, and worth retaining under human judgment.

The key point is not simply that theory-oriented social science differs from mathematics, but that this difference forces a redefinition of search, verification, branching, and success. Therefore, AI assistance for theory-oriented social science should be evaluated and designed around assumption search, human adjudication, and research-object quality, rather than around proof-centric success criteria alone.

Therefore, proof-centric benchmarks are incomplete for this setting: they may capture local reasoning quality, but they miss whether a system leaves behind an inspectable research object worth retaining under human judgment.

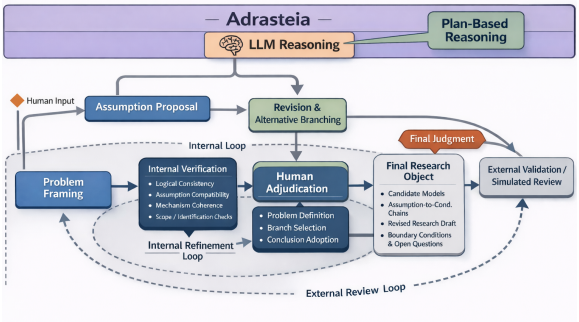

Figure 1. **Illustrative workflow instantiation.** Inner proposal–verification loop, explicit human adjudication, and outer simulated review.

4. A Concrete Instantiation of the Position

If the preceding argument is right, then the next question is architectural: what would a system look like if it were organized around assumption search rather than proof search? To make the position concrete, we use ADRASTEIA as an illustrative workflow instantiation. The point of this section is not to claim that this is the only correct system design, but to show what follows if AI assistance for theory-oriented social science is organized around problem framing, assumption proposal, diagnostic verification, human adjudication, and external review. Figure 1 summarizes the resulting structure: an inner loop of proposal, verification, branching, and human adjudication, together with an outer loop of reviewer-style critique and return.

4.1. Problem Framing

If the starting point is a phenomenon, mechanism intuition, or target claim rather than a fixed proposition, then system design must begin with problem framing rather than direct derivation. In economics-centered theory work, research often begins not from a theorem statement that has already been cleanly specified, but from a phenomenon to be explained, a mechanism to be explored, or a claim that the researcher wishes to discipline theoretically.

The role of problem framing is therefore to transform an initially vague research concern into a structured object of inquiry. This includes identifying what is to be explained, what class of mechanisms is under consideration, what kinds of assumptions may become relevant, and what would count as a satisfactory theoretical outcome. Different reasoning backends may instantiate this stage, but the architectural point is the stage itself rather than any single model.

4.2. Assumption Proposal and Branching

If the bottleneck is assumption search, then the workflow must search over candidate model worlds rather than over derivation paths alone. Instead of generating only answers

or proofs, it should propose candidate modeling structures: primitives, information assumptions, institutional constraints, equilibrium environments, and alternative mechanisms.

The output is therefore not a single model, but a set of candidate assumption packages and mechanism branches. Branching is not an after-the-fact recovery step. It is part of the search logic itself, because multiple assumption packages may remain viable and intellectually meaningful for long stretches of the research process.

4.3. Verification under Weak External Validation

If theory-oriented research is evaluated under weak and heterogeneous external standards, then verification cannot be reduced to formal accept/reject checking. It must instead diagnose where assumptions are too strong, where mechanism claims are under-motivated, where citations or factual background are misrepresented, and where prose creates an illusion of rigor without genuine inferential support.

In ADRASTEIA, this combines literature-grounding and factual verification with proof- and text-level hallucination checks. The point is to flag overstated assumptions, hidden jumps, variable drift, and claim overreach. Verification is therefore not a terminal certificate of correctness, but structured feedback that guides revision, comparative branching, and judgment.

4.4. Human Adjudication

If assumption acceptability is partly disciplinary and interpretive, then human adjudication must be treated as a first-class architectural component rather than as fallback supervision. Human intervention is not a patch for model failure. It is the mechanism through which originality, interpretive control, and scholarly accountability are preserved in settings where formal validity alone does not determine research value.

In practice, the researcher intervenes at problem definition, candidate assumption selection, interpretation of verifier outputs, branch retention, and conclusion framing. Those interventions determine which branches count as meaningful progress and which conclusions are disciplined enough to keep.

4.5. Research Objects Rather than Final Answers

If the success criterion is not "proof accepted" but "research object worth retaining," then the workflow should aim to produce an inspectable research object rather than a final answer. In ADRASTEIA, that object includes candidate models, explicit assumption-to-conclusion chains, revised and verifier-audited drafts, and records of boundary conditions, unresolved issues, and open alternatives.

The aim is not to terminate the research process with a single conclusion, but to leave behind an object that can be inspected, criticized, extended, and reviewed. The workflow therefore makes the conclusion, its assumptions, its alternatives, and its limitations visible at the same time.

4.6. External Review as a Design Consequence

If theory-oriented research is judged under weak external validation, then simulated review becomes a natural design consequence rather than an optional add-on. The workflow therefore does not end once a research object has been constructed. It should also expose that object to reviewer-style criticism before further iteration.

The purpose of this stage is not to replace human referees, but to surface additional objections before another round of revision. Review-oriented systems such as Refine or Stanford Agentic Reviewer can supply this pressure test, but they are not ground truth; their role is to feed new concerns back into revision and human adjudication. The result is a coupled inner loop of proposal, verification, branching, and judgment, together with an outer loop of critique and return.

5. Alternative Views and Scope Conditions

5.1. Alternative Views

Alternative view: some theory-oriented research remains close to formal reasoning. Not all theory-oriented social science is equally far from mathematics. Some projects do reduce to tightly specified formal reasoning problems, and in those cases proof-centric systems remain valuable. Our claim is therefore not that proof search is useless, but that it is the wrong default abstraction for the broader class of problems in which assumptions, scope conditions, and interpretations remain open.

Alternative view: stronger general agents may remove the need for explicit assumption-search framing. One might argue that sufficiently strong general research agents will absorb this difficulty automatically. That possibility should not be dismissed. But greater capability does not by itself solve a problem of task specification. If the search object and success criterion are mischaracterized, stronger agents may simply produce more fluent but less inspectable outputs. The argument of this paper is therefore about problem formulation as much as raw capability.

Alternative view: assumption search is only workflow engineering by another name. Another objection is that "assumption search" merely redescribes a workflow preference rather than identifying a distinct epistemic structure. Our view is that the workflow consequence is downstream

of a deeper difference: the object being searched, the semantics of verification, and the criterion of success are all different from the proof-centric case. Workflow engineering matters here because it operationalizes that difference; it is not the whole claim.

5.2. Scope Conditions

Economics-centered scope. Although the paper is framed at the level of theory-oriented social science, its primary arguments, examples, and empirical illustrations are centered on economics. Any extension beyond economics should therefore be read as cautious extrapolation to adjacent theory-oriented settings rather than as a claim of uniform coverage across the social sciences.

No unified formal checker. Theory-oriented social science lacks a universal checker analogous to a formal proof assistant. Verification is therefore necessarily weaker, more heterogeneous, and more interpretive than proof verification. This is a scope condition of the domain, not merely a defect of one implementation.

Assumption acceptability is difficult to formalize. Naturalness, parsimony, plausibility, and disciplinary acceptability are only partly articulable and depend on field history and scholarly taste as well as logic. AI systems can make assumptions more explicit and comparable, but they cannot remove the judgment-laden character of assumption selection itself.

Human adjudication remains indispensable. Human intervention is not merely a fallback. In the class of problems discussed here, final quality depends partly on the researcher's expertise, taste, and goals. That dependence is not an embarrassment for the position; it is one of the reasons proof-centric automation does not transfer cleanly.

Open problem: from workflow assistance to durable research practice. A final open question is how systems such as ADRASTEIA fit into longer-run norms of originality, authorship, and credit. The challenge is institutional as well as technical: increasing rigor and productivity without eroding judgment and scholarly responsibility.

Taken together, these alternative views and scope conditions do not overturn the position advanced here. They narrow it. The claim is not that proof-centric AI has no role, nor that one workflow settles the question, but that a large and important part of theory-oriented social science is better understood as assumption search under human judgment.

6. Illustrative Empirical Support

The empirical material in this paper comes from an IRB-approved evaluation program that is still in progress. We therefore report only the partial results needed to motivate the position here. After the evaluation is complete, we will release the full experimental protocol, prompts, robustness checks, and extended tables.

The current results are encouraging in two respects. First, in a compute-matched automated reviewer-style proxy comparison, Automated ADRASTEIA reaches a 58.7% task-level win rate against the self-reflection baseline. Second, the gains are not evenly distributed: they are smallest in comparative statics and clearest in assumption-search and revision-heavy tasks. This concentration is the relevant pattern for our argument. It suggests that the workflow's value lies less in dominating every local derivation step than in supporting assumption-heavy and critique-heavy research operations.

These results remain preliminary. Automated review is a proxy rather than ground truth, realized performance depends on backend and reviewer configuration, and the benchmark covers only a limited slice of theory-oriented research. Still, the partial evidence is consistent with the position's main prediction: assumption-search workflows are most useful where research progress depends on framing, critique, and revision under human judgment.

7. Conclusion

We argued that AI-for-math paradigms should not be treated as the default template for AI-assisted theory-oriented social science. In these settings, the central bottleneck is often assumption search under weak external validation, not proof search under strong verification. That difference changes the search object, the semantics of verification, the role of human judgment, and the criterion of success for AI assistance.

ADRASTEIA is offered as one concrete instantiation of that position rather than as the paper's main claim. The empirical evidence reported here is illustrative and suggests that assumption-heavy and revision-heavy tasks are where this framing is most useful, but it does not adjudicate the position in full. The practical implication for the field is therefore straightforward: future AI systems for theory-oriented social science should be designed and evaluated around assumption search, human adjudication, diagnostic verification, and research-object quality, rather than around proof-centric success criteria alone.

References

[1] Feng, T., Trinh, T. H., Bingham, G., et al. Towards Autonomous Mathematics Research. *arXiv preprint*

arXiv:2602.10177, 2026.

[2] Gottweis, J., Weng, W.-H., Daryin, A., et al. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

[3] Achim, T., Best, A., Der, K., et al. Aristotle: IMO-level Automated Theorem Proving. *arXiv preprint arXiv:2510.01346*, 2025.

[4] Hubert, T., Mehta, R., Sartran, L., et al. Olympiad-level Formal Mathematical Reasoning with Reinforcement Learning. *Nature*, 651, 607–613, 2026.

[5] Trinh, T. H., Wu, Y., Le, Q. V., He, H., and Luong, T. Solving Olympiad Geometry without Human Demonstrations. *Nature*, 625, 476–482, 2024.

[6] Lean Language Team. Learn. Lean documentation website. `https://lean-lang.org/learn/`. Accessed 2026-04-21.

[7] Romera-Paredes, B., Landajuela, M., Ha, D., et al. Mathematical Discoveries from Program Search with Large Language Models. *Nature*, 625, 2024.

[8] Woodruff, D. P., Cohen-Addad, V., Jain, L., et al. Accelerating Scientific Research with Gemini: Case Studies and Common Techniques. *arXiv preprint arXiv:2602.03837*, 2026.

[9] Feng, T., Trinh, T., Bingham, G., et al. Semi-Autonomous Mathematics Discovery with Gemini: A Case Study on the Erdős Problems. *arXiv preprint arXiv:2601.22401*, 2026.

[10] Korinek, A. Language Models and Cognitive Automation for Economic Research. *NBER Working Paper No. 30957*, 2023.

[11] Korinek, A. Generative AI for Economic Research: Use Cases and Implications for Economists. *Journal of Economic Literature*, 61(4), 2023.

[12] Korinek, A. AI Agents for Economic Research. *NBER Working Paper No. 34202*, 2025.

[13] Ludwig, J., Mullainathan, S., and Rambachan, A. Large Language Models: An Applied Econometric Framework. *NBER Working Paper No. 33344*, 2025.

[14] Horton, J. J., Filippas, A., and Manning, B. S. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? *NBER Working Paper No. 31122*, 2023.

[15] Dawid, H., Harting, P., Wang, H., Wang, Z., and Yi, J. Agentic Workflows for Economic Research: Design and Implementation. *arXiv preprint arXiv:2504.09736*, 2025.

[16] Feyzollahi, M. and Rafizadeh, N. The Adoption of Large Language Models in Economics Research. *Economics Letters*, 250:112265, 2025.

[17] Guo, Y. and Yang, Y. EconNLI: Evaluating Large Language Models on Economics Reasoning. *Findings of the Association for Computational Linguistics: ACL 2024*, pages 982–994, 2024.

[18] Wu, J. C., Xi, J., and Xie, S. LLM Survey Framework: Coverage, Reasoning, Dynamics, Identification. *NBER Working Paper No. 34308*, 2025.

[19] Steinberg, J. Trade Policy Ambiguity. *NBER Working Paper No. 34408*, 2025.

[20] Rein, D., Hou, B. L., Stickland, A. C., et al. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *Proceedings of COLM*, 2024.

[21] Hendrycks, D., Burns, C., Kadavath, S., et al. Measuring Mathematical Problem Solving With the MATH Dataset. *NeurIPS*, 2021.

[22] Autonomous Policy Evaluation Project. About APE. `https://ape.socialcatalystlab.org/about`

[23] Autonomous Policy Evaluation Project. Methodology. `https://ape.socialcatalystlab.org/methodology`

[24] Autonomous Policy Evaluation Project. Project History and Tournament Description. `https://ape.socialcatalystlab.org/history`

[25] Autonomous Policy Evaluation Project. Judge Calibration. `https://ape.socialcatalystlab.org/calibration`