Discuss this article at JournalTalk:
https://journaltalk.net/articles/n6232/

EJW
ECON JOURNAL WATCH
Scholarly Comments on Academic Economics

ECON JOURNAL WATCH 23(2)
September 2026: 509–522

Are Chatbot Validations of Mathematical Economics Valid?

Alexis Akira Toda[1]

LINK TO ABSTRACT

Over the years, I have been involved in exposing or correcting errors in four papers—including one of my own—published in leading economics journals:

1. Tirole (1985): Pham and Toda (2026) show that Proposition 1(c) is mathematically incorrect by presenting a counterexample, and they restore the result under additional assumptions;[2]

2. Kocherlakota (1992): the original proof of Proposition 4 was incorrect, but Kocherlakota and Toda (2023) corrected it under the original assumptions;

3. Miao and Wang (2018): the paper defined "bubble" inconsistently, and Hirano and Toda (2025b) proved the nonexistence of bubbles in their model according to the standard definition;

4. Stachurski and Toda (2019): the proof of Proposition 5 was incorrect, but Stachurski and Toda (2020) corrected it by strengthening the assumptions (in an economically harmless way).

After the publication of Pham and Toda (2026), several people asked me if we used artificial intelligence (AI) to find the error or write the proof. The answer is no—we found the error and wrote the proof on our own, although we did use AI for grammar checking, as neither of us is a native English speaker. But the

[1] Department of Economics, Emory University. Email: alexis.akira.toda@emory.edu. I thank Emi Nakamura for helpful comments and suggestions, and I gratefully acknowledge financial support from the Zengin Foundation.

[2] It may come as a surprise that a highly cited paper written by a Nobel laureate escaped scrutiny for more than 40 years. As of May 2026, Tirole (1985) had more than 2,000 Google Scholar citations.

question made me curious. The mathematical ability of AI models has improved remarkably fast; OpenAI even announced that one of its large language models found a solution to the Navier-Stokes existence and smoothness problem.[3] So I wanted to see: if I fed these four papers to chatbots and asked whether the results are correct, would they catch the errors I know are there?

The experiment was simple. For each paper, I uploaded the document to several chatbots (Gemini, Refine, Claude, and ChatGPT), asked whether the key result is correct, and then challenged the model iteratively, steering it toward the problematic part as needed. Since I know each flaw and its resolution intimately, I am well positioned both to prompt the models and to judge their answers. In particular, I wanted to see two things: first, whether the chatbots would generate *false positives*, that is, endorse a flawed proof as correct; and second, how much guidance a chatbot needs before it finds the error.

The simple results are that the chatbots did generate false positives, and that they varied in how much guidance was needed. On the first pass at Tirole (1985), every model told me the proof was correct, or essentially so, and the flaw surfaced only after I steered the conversation toward it. ChatGPT Pro was the clear winner: in several cases it flagged the relevant issue from a single prompt and constructed valid counterexamples and corrected proofs. Claude was weaker on formal reasoning but stronger on economic interpretation. Gemini performed worst.

One caveat is worth flagging up front. A claim that a model "found" an error must be read against its *knowledge cutoff*, the date beyond which it has not been trained on new data. Most of the papers I study and their corrections have circulated for some time, so a model may simply have seen the relevant material during training rather than reasoned it out independently. To guard against this, I took particular care with Tirole (1985), whose correction was published only in May 2026 in Pham and Toda (2026).[4] Furthermore, for all experiments with ChatGPT Pro (and only that chatbot), I disabled both the memory and the web-search functions before running the analysis.

Next, I will explain the experiments paper by paper.[5] Afterward, I say what I make of it all.

[3] https://openai.com/index/navier-stokes-solution/

[4] However, their working paper version (https://arxiv.org/abs/2507.12477) was posted in July 2025.

[5] Links to the full chat transcripts are in the footnotes; in case those links no longer work, I have also saved all the transcripts as PDFs and posted them as supplementary material.

The Case of Tirole (1985)

Again, Proposition 1(c) of Tirole (1985) is incorrect. As I am not very tech-savvy, I did not know how to use AI to check mathematical correctness (as of May 26, 2026). As a first step, I asked Google Gemini 3.5 Flash (which is free and built into the Google Chrome browser):

> I want to check the mathematical correctness of an economics paper. The paper is in a PDF format. Can you explain to me how to do it using Gemini (or other free AI tools)?

Gemini gave me some instructions and I followed them.[6]

Gemini 3.5 Flash

After the initial prompts, I directly asked about the proposition in question:

> Is the proof of Proposition 1(c) logically correct?

To this prompt, Gemini replied that "The proof of Proposition 1(c) in Tirole's (1985) paper is logically sound" and explained the key steps, including cases (i)–(iii) for the asymptotic behavior of the interest rate. Although I knew case (ii) was problematic, instead of flagging this case specifically, I asked for more clarification:

> In the paper, the possibility of a bubbleless equilibrium with $\bar{r} < 0$ is considered at the bottom of page 1522, where Tirole states 'Let us now show that if $\bar{r} < 0$, there exists no [bubbleless] equilibrium.' As you mentioned, Tirole considers three exhaustive and mutually exclusive cases. In each case, as you mention, Tirole reasons that $f_t \to 0$ and hence $r_t \to \bar{r}$. But I don't understand why $f_t \to 0$ implies $r_t \to \bar{r}$.

Gemini replied that "your skepticism is completely justified. The leap from $f_t \to 0$ to $r_t \to \bar{r}$ isn't immediately obvious from the text, but it relies on the continuity and monotonic properties of the temporary equilibrium mapping" and provided some more details. However, Gemini's analysis was flawed because it claimed global stability from local stability, so I further challenged it:

> Thank you for the clarification, but I still don't understand the argument. On p. 1502 of the Tirole paper, he states 'We also make Diamond's stability assumption that the curves defined by $\{w = \Phi(r)\}$ and $\{r = \psi(w, 0)\}$ have a unique intersection and that the ψ curve is strictly steeper (in absolute value) than the Φ curve at this intersection in the (r, w) plane.' This is a local argument, but

[6] https://gemini.google.com/share/5eaafdf936f1

> your explanation seems to rely on a global argument. In general, a local stability condition guarantees that the dynamical system converges near the steady state, but there is no guarantee when the state is far away.

Gemini admitted that "You are completely correct that Diamond's stability assumption, as stated on page 1502, is a local stability condition," but then claimed that global convergence can be obtained due to the monotonicity property in the Diamond model. However, this reasoning was flawed, so I challenged further:

> Thank you for the clarification, but I think there is still an issue. If the OLG model does not have dividends (what Tirole calls rents), then the dynamics is one-dimensional, so the uniqueness of the steady state and the monotonicity property guarantee global convergence. But with positive dividends (rents), the dynamics is two-dimensional, so I don't think the proof works as you suggest. Am I missing something?

To this prompt, Gemini explained the proof of Proposition 1(c) step-by-step. I then directly flagged the leap of logic:

> I am still uncomfortable with Tirole's (and your) logic. In the logic flow you describe, I agree until Step 3. But in Step 4, Tirole (or you) claim that because $f_t \to 0$, the system behaves asymptotically like the one-dimensional Diamond model, forcing $r_t \to \bar{r}$. I think there is a leap of logic here. Consider Case (ii), where $\{r_t\}$ is strictly greater than n. Then we know that equilibrium forces $f_t \to 0$. But if $r_t \to \infty$ simultaneously, capital k_t will converge to 0. Such a possibility does not seem to be ruled out by Tirole's argument.

To my challenge, Gemini basically claimed that the Inada condition $f'(0) = \infty$ prevents equilibrium paths converging to zero. I knew this argument was flawed, so I simply uploaded Pham and Toda (2026), after which Gemini admitted its mistake.

Overall, after each prompt, Gemini 3.5 Flash provided a plausible argument, but it was not well-founded.

Refine

Next, I turned to Refine.[7] The reason I used this tool is that, when Pham and Toda (2026) was conditionally accepted at *Econometrica*, the editor ran our paper through Refine and asked us to provide feedback on the report it generated. (The Refine report was sophisticated but contained incorrect statements.)

Refine's service costs $50 per analysis. Upon uploading Tirole (1985), Refine flagged several issues, but none of them looked serious; I found the report nitpicking. However, I used the chat function to discuss further.[8] After several exchanges

[7] https://www.refine.ink/
[8] https://www.refine.ink/app/session/17a36e9c-5189-441b-a96e-b55ce478467a

similar to those with Gemini, Refine admitted that the proof of Proposition 1(c) was incomplete. Furthermore, when I asked it to construct a counterexample (with few hints), it successfully generated a counterexample based on the Cobb-Douglas production function, though several details were omitted.

Overall, although the initial Refine report was somewhat underwhelming, its chat function was impressive. However, I am not sure if I would regularly use it from a cost perspective.

Claude Opus 4.8

Next, I turned to Anthropic's Claude Opus 4.8 Pro, which costs $20 per month. I did not use the more expensive ($100) Claude Opus 4.8 Max because the pricing information suggested that Max allows for more volume and speed, not additional functionality.[9] My discussion with Claude was similar to Gemini: initially, Claude claimed that Proposition 1(c) was correct, but after I challenged it a few times, it admitted that there was a leap of logic.[10] When I asked Claude to generate a counterexample, it tried the Cobb-Douglas function and derived necessary conditions for what the counterexample would look like, but it failed to generate one. When I provided the functional form

$$f(k) = Ak \log(1 + 1/k) \tag{1}$$

used in the counterexample in Pham and Toda (2026), Claude solved the model numerically and claimed that it obtained a counterexample. However, upon inspection, it did not satisfy one of the assumptions of Proposition 1(c), so the analysis was incomplete.

Overall, I was impressed with Claude's computational capabilities, but not necessarily with its mathematical reasoning.

ChatGPT

Finally, I turned to OpenAI's ChatGPT. First, I tried the free version.[11] It made some basic mathematical mistakes, for example, mixing up the convergence of a subsequence with that of a sequence. (See the part titled "6. The precise hidden argument," where it mistakenly concludes $r_{t_k+1} \to r^*$ after establishing $r_{t_k} \to r^*$.)

Next, I upgraded to ChatGPT 5.5 Pro, which costs $100 per month. Before running the experiment, I was careful to turn off ChatGPT's memory and the web-

[9] https://claude.com/pricing. When I ran the experiment on May 30, 2026, the more sophisticated model Claude Fable 5 was not yet available.
[10] https://claude.ai/share/255203c1-1de8-44d3-9536-cd8575caf2bd
[11] https://chatgpt.com/share/6a18d3ff-a5dc-83ea-80ea-0b717740a47a

search functions. ChatGPT Pro's analysis was truly amazing.[12] Even though I only asked "Can you check the mathematical correctness of the proof of Proposition 1(c)?", ChatGPT Pro flagged several issues and replied "correct result, correct core logic, but not a fully rigorous proof as printed unless one supplies the missing boundary/supremum argument in the $\bar{r} < 0$ existence step." I then pointed out the key issue:

> You mentioned 'The no-bubbleless-equilibrium part is essentially correct' but I have an issue here. You mentioned 'If $f_t \to 0$, then the economy asymptotically behaves like Diamond's rentless/bubbleless economy, so $r_t \to \bar{r} < 0$.' In the paper, the possibility of a bubbleless equilibrium with $\bar{r} < 0$ is considered at the bottom of page 1522, where Tirole states 'Let us now show that if $\bar{r} < 0$, there exists no [bubbleless] equilibrium.' Tirole considers three exhaustive and mutually exclusive cases. In each case, Tirole reasons that $f_t \to 0$ and hence $r_t \to \bar{r}$. But I don't understand why $f_t \to 0$ implies $r_t \to \bar{r}$.

With only this prompt, ChatGPT Pro (correctly) raised an issue with case (ii) discussed with Gemini and (correctly) concluded that "this case does not itself give monotonicity […] the sentence '$f_t \to 0$, hence $r_t \to \bar{r}$' is not mathematically justified as a standalone implication." Upon asking for a counterexample, ChatGPT Pro tried the case with a utility function exhibiting constant relative risk aversion (CRRA) $\gamma > 1$ and the production function

$$f(k) = \frac{A}{\gamma} k[\log(1/k)]^{\gamma}. \tag{2}$$

This suggestion was truly remarkable, as the special case of (2) with $\gamma = 1$ (Cobb-Douglas) becomes $f(k) = Ak\log(1/k)$, which behaves exactly the same way as the function (1) as $k \to 0$. After several prompts, ChatGPT Pro was able to provide a counterexample with a complete proof.

Overall, I was hugely impressed with the mathematical ability of ChatGPT Pro.

The Other Three Papers

After my experiment with Tirole (1985), to test the robustness of my findings, I decided to feed the other three papers into the chatbots. Before running the experiment, I set the following rules:

1. The papers I will test are Kocherlakota (1992), Miao and Wang (2018), and Stachurski and Toda (2019).

[12] https://chatgpt.com/share/6a1a519e-e57c-832d-9b82-53816c2b5946

2. The AI models I will test are Gemini 3.5 Flash, Claude Opus 4.8, and ChatGPT 5.5 Pro, with memory and web-search turned off for the latter until instructed otherwise.[13]

3. For each paper, my initial prompt would be to check the mathematical correctness of the key result. After the initial prompt, I will challenge as needed.

The case of Kocherlakota (1992)

For each AI model, the initial prompt was as follows:

> I have just uploaded the paper 'Bubbles and Constraints on Debt Accumulation' published in Journal of Economic Theory. Can you check the mathematical correctness of the proof of Proposition 4?

Gemini 3.5 Flash Initially, Gemini responded "The logic presented in the proof is mathematically sound."[14] Upon challenging, Gemini (incorrectly) claimed that the existence of a subsequence $\{t_n\}$ and a positive constant $b > 0$ such that $a_{t_n} - a_{t_n-1} > b$ is a logical consequence of the assumption that $\{a_t\}$ does not converge. When I further challenged Gemini by pointing out that there could be a sequence such that $\{a_t\}$ does not converge but $a_t - a_{t-1}$ converges to zero, Gemini admitted its mistake and provided the (correct) counterexample $a_t = \sin(\sqrt{t})$. When I asked if there is a known correction to Kocherlakota (1992), Gemini was unable to find Kocherlakota and Toda (2023). When I provided the exact link of the corrigendum, Gemini hallucinated.

Claude Opus 4.8 Initially, Claude responded "The proof of Proposition 4 is essentially correct in its logic, but it relies on a step that is asserted rather than fully justified, and there's a notational slip."[15] However, Step 2 contained a flaw. Upon challenging, Claude provided a counterexample very similar to the one in Section 3 of Kocherlakota and Toda (2023), as well as $a_t = \sin(\sqrt{t})$ (same as Gemini). When I asked if there is a known correction to Kocherlakota (1992), Claude was able to find Kocherlakota and Toda (2023).

---

[13] I exclude Refine because I did not want to pay $50 each time, and it was no better than ChatGPT Pro based on the previous experiment. For chatbots other than ChatGPT Pro, I did not know how to turn off memory and web-search.

[14] https://gemini.google.com/share/1c05411055f6

[15] https://claude.ai/share/5d4768af-2d65-4f31-bf3a-c5cbcbc9dd5a

ChatGPT 5.5 Pro With just my initial prompt, ChatGPT Pro identified the error in the proof of Proposition 4, provided the (correct) counterexample $a_t = \sin(\log(t+1))$, and furthermore, correctly proved the original statement, claiming that "A corrected proof is straightforward."[16] In fact, ChatGPT Pro's proof is more elegant than Kocherlakota and Toda (2023)! When I asked if there is a known correction to Kocherlakota (1992), ChatGPT Pro was able to find Kocherlakota and Toda (2023).

The case of Miao and Wang (2018)

> For each AI model, the initial prompt was as follows:
>
> I have just uploaded the paper 'Asset Bubbles and Credit Constraints' and its online appendix. The abstract claims 'We provide a theory of rational stock price bubbles in production economies with infinitely-lived agents.' The authors claim on p. 2595 that 'Some studies (e.g., Scheinkman and Weiss, 1986; Kocherlakota, 1992, 2008; Santos and Woodford, 1997; Hellwig and Lorenzoni, 2009) have found that infinite-horizon models of endowment economies with borrowing constraints can generate rational bubbles. Unlike this literature, our paper analyzes a production economy with stock price bubbles attached to productive firms.' The papers cited here are all 'rational bubble' models, where the asset price exceeds the fundamental value defined by the present discounted value of dividends. Now, based on the standard definition of rational bubbles, does the model of Miao and Wang generate bubbles?

Gemini 3.5 Flash Initially, Gemini (incorrectly) responded "Based on the standard definitions and the specific modeling choices presented in the paper, the model of Miao and Wang does generate rational asset bubbles."[17] In the Miao-Wang model, the stock price takes the form $V_t = Q_t K_t + B_t$. Gemini claimed that $Q_t K_t$ is the fundamental value and B_t is the bubble. When I asked if the authors prove that $Q_t K_t$ is the fundamental value, Gemini said yes, and cited the transversality condition. When I pointed out a contradiction, Gemini found my website, on which I discuss these issues, as well as my published article (Hirano and Toda 2025b) and other papers.

Claude Opus 4.8 Claude (correctly) responded "Based on the standard definition of a rational bubble, the answer is no …the honest characterization is the following. The model generates multiplicity of equilibria—a 'bubbly' and a 'bubbleless' steady state—driven by self-fulfilling beliefs […]. But labeling the high-value equilibrium's extra term a 'rational bubble' in the Santos-Woodford sense is

[16] https://chatgpt.com/share/6a1cc5bc-0c58-8333-82cf-403872e82c12
[17] https://gemini.google.com/share/ca5c671327d8

a strain," and "Whether that broadened notion deserves the name 'rational bubble' is exactly the point of contention, and reasonable people in the field disagree."[18] Upon being asked to elaborate on "reasonable people in the field disagree," Claude conceded that it was a speculation, but then searched the web and (correctly) summarized the exchange of Hirano and Toda (2025b), Miao and Wang (2025), and Hirano and Toda (2025a).

ChatGPT 5.5 Pro Initially, ChatGPT Pro (correctly) responded "no, not in the usual Santos-Woodford/asset-pricing sense" and "Calling that object a 'bubble' is internally consistent with their terminology, but it is not the same object as the rational bubbles in the papers they cite."[19] When I asked for clarity about two different notions of "transversality condition" (TVC) and whether Miao and Wang actually proved it, ChatGPT (correctly) answered "They do not prove Equation (6) as an independent household optimality/no-Ponzi TVC inside the paper. They mostly state it, cite general references, and then use it." When I pointed out language issues in the abstract and introduction, ChatGPT Pro provided the verdict "This makes the abstract and literature framing problematic. When they contrast their paper with Scheinkman-Weiss, Kocherlakota, Santos-Woodford, Hellwig-Lorenzoni, and related 'rational bubble' models, they naturally invite the reader to understand their result in the same rational-bubble sense. […] They market this mechanism as a rational stock-price bubble, even though, under the standard Santos-Woodford present-value definition, their imposed TVC rules out such bubbles." Once I enabled web search, ChatGPT Pro found the exchange of Hirano and Toda (2025b), Miao and Wang (2025), and Hirano and Toda (2025a) and concluded that Hirano and Toda's "critique is essentially the same as the one we have been discussing" (i.e., the analysis of ChatGPT Pro).

The case of Stachurski and Toda (2019)

> For each AI model, the initial prompt was as follows:
>
> I have just uploaded the paper "An impossibility theorem for wealth in heterogeneous-agent models with limited heterogeneity" published in Journal of Economic Theory. Can you check the mathematical correctness of the proof of Proposition 5?

Gemini 3.1 Pro Being disappointed with Gemini 3.5 Flash, I broke my own rule and switched to Gemini 3.1 Pro, which is free but supposed to be stronger at math and code. Gemini initially responded "Based on a careful review of the provided

[18] https://claude.ai/share/65098d32-6af6-46d3-9760-92a5889733d8
[19] https://chatgpt.com/share/6a1d011e-a504-8331-b445-6fd18b2d6ff8

paper, the mathematical proof of Proposition 5 is rigorous and logically sound."[20] When I challenged Gemini about continuity issues, it admitted that the proof had a gap and suggested ideas to fix it. It also found the correction by Stachurski and Toda (2020).

Claude Opus 4.8 After a lengthy analysis, Claude responded "The proof of Proposition 5 is mathematically correct."[21] Upon my challenging it on issues on continuity, Claude admitted the gap but claimed the proof can be rescued. When asked if a correction exists, it found Stachurski and Toda (2020), though it was unable to access the paper. When I uploaded the corrigendum, Claude flagged that the original proof had a further issue (pointed out in the corrigendum).

ChatGPT 5.5 Pro Just with my initial prompt, ChatGPT identified key gaps in the proof of Proposition 5 and concluded "the proposition is very likely true, but the proof as written is not fully mathematically correct."[22] When asked if a correction exists, it found Stachurski and Toda (2020) and confirmed that the proof of their Proposition 5 is correct.

What I Make of It

False positives and the guidance needed

The headline result is easy to state. When simply handed a paper and asked whether the key result is correct, the chatbots generated false positives, and endorsed flawed proofs. This was clearest for Tirole (1985), where every model's first response was that the proof was correct or essentially so. The error came to light only because I knew in advance which part was problematic, and I steered the conversation accordingly. A reader who simply uploaded the paper and asked whether the proof was correct would have been told yes.

How much guidance was needed varied considerably across models, and this is where the model-by-model comparison is informative. ChatGPT Pro was the clear winner. It was the only model that, with no more than my initial prompt, repeatedly flagged the relevant issues, and in two cases (Kocherlakota 1992, and the existence step in Tirole 1985) it constructed a valid counterexample and a more elegant corrected proof. Claude Opus 4.8 was less mathematically sophisticated, tending to assert rather than justify the crucial step, but it was stronger on judgment and interpretation. In the case of Miao and Wang (2018), for instance, it correctly

[20] https://gemini.google.com/share/0273b68c747d
[21] https://claude.ai/share/db71e110-e439-45c1-9f2c-6aebbf1edc88
[22] https://chatgpt.com/share/6a1d1879-0508-8333-b4d0-89c56f7d25fa

resisted the "rational bubble" framing and characterized the disagreement in the field. Gemini was the clear loser. Across papers, it initially endorsed incorrect proofs, defended them with plausible but unfounded arguments, and on one occasion hallucinated when asked for a specific reference.

Even for the best model, the performance should not be read as a claim that it is an autonomous error-finder. This point is easy to miss if one only skims the chat transcripts. The chatbots are excellent at following a line of reasoning once it is pointed out, at performing the algebra, and, in the case of ChatGPT Pro, at completing a proof whose key idea has been identified. But they did not, on their own, decide that some results in the papers were incorrect. The practical implication for an economist is the following. Running a paper through even the best available AI model and asking whether the result is correct will probably not surface a deep error without substantial input from a human who already suspects where the problem lies. The chatbot lowers the cost of *checking* an argument once a candidate flaw is identified. It does not yet reliably substitute for the human judgment that *locates* the flaw in the first place.

What this suggests about peer review

It is natural to ask how this human-plus-chatbot benchmark compares with peer review as actually practiced. My experience with the four papers is not encouraging on that score. Each of them contains a flawed result that survived refereeing at a top journal, though it was eventually corrected or criticized. The failures run in the other direction too: when Hirano and Toda (2025b) was submitted to *American Economic Review*, the original authors recommended rejection, and some of the three independent referees dismissed the issue as "merely" semantic. Based on their reports, none of them seemed to have read the proofs. Thus, the current system both accepts incorrect results and rejects correct critiques.

Part of the reason, I suspect, is that few referees read proofs in mathematical economics.[23] Refereeing is essentially unpaid work, so even well-meaning referees economize on effort. Against this backdrop, my tentative view is that for evaluating the internal logic of mathematical economics, a decently competent person using a frontier model such as ChatGPT Pro will likely outperform the typical referee. The emphasis belongs on *a person using AI*: the human supplies domain knowledge and targeted skepticism while the model supplies patient checking, algebra, and

[23] As an example, consider Stachurski and Toda (2019). As I document in my blog post (https://alexisakira.github.io/publications/2019-JET/), this paper was rejected from *American Economic Review: Insights* (three reports), *Econometrica* (desk rejection upon consulting an external advisor), and *Review of Economic Studies* (four reports plus an external advisor), and accepted as is at *Journal of Economic Theory* (one report). Among 10 reviewers/advisors who evaluated the paper, none seemed to have read the proof.

counterexample search. This has an implication for editorial policy. Although some journals prohibit reviewers from uploading submitted papers to AI models, such a prohibition is difficult to enforce and, in my view, already out of date. The legitimate concern behind it is confidentiality, since a chatbot provider could in principle train on the uploaded text. But this concern is easily sidestepped: journals could simply ask authors to post their papers on a public repository such as arXiv before submission, after which any confidentiality would have already been waived.

Caveats

A few caveats are in order:

1. The exercise is based on a small set of papers.

2. Each paper was selected because I happened to know an error in it. This is a feature for prompting quality but a limitation for external validity, and my role as an author of several of the corrections may color the interpretation.

3. The comparison across models is not fully controlled. Pricing tiers, default settings, and tool access differ, and the models are moving targets that may behave differently a few months from the point of the experiments (May 2026).

4. As discussed above, I cannot fully exclude the possibility that a model recognized a paper and retrieved the known correction from its training data rather than immanently reasoning to it, since most of the papers and their corrections predate the models' knowledge cutoffs. The Tirole (1985) experiment, run with memory and web-search disabled and almost concurrently with the publication of the correction, is the cleanest case on this score—and, again, the result was a false positive, at least initially. For the remaining papers the picture is more mixed. That said, the counterexample ChatGPT Pro produced for Tirole (1985) was different from the one in Pham and Toda (2026), which makes me reasonably confident it was genuinely constructed rather than retrieved.

Taken together, the experiments suggest cautious optimism. Frontier AI is already a useful aid for scrutinizing mathematical economics, and a knowledgeable human working with it is plausibly better than the current practice of referee reports written by humans who do not read proofs (or who have conflicts of interest). But can chatbots validate mathematical economics on their own? Not yet: left to themselves, they say "correct" too easily.

Supplementary Material

Transcripts of the chatbot conversations can be downloaded here.

References

Hirano, Tomohiro and Alexis Akira Toda. 2025a. Cry Wine and Sell Vinegar: Rejoinder to Miao and Wang. *Econ Journal Watch* 22(2): 211–231. Link

Hirano, Tomohiro and Alexis Akira Toda. 2025b. Toward Bubble Clarity: A Comment on Miao and Wang. *Econ Journal Watch* 22(1): 1–17. Link

Kocherlakota, Narayana. 1992. Bubbles and Constraints on Debt Accumulation. *Journal of Economic Theory* 57(1): 245–256. Link

Kocherlakota, Narayana and Alexis Akira Toda. 2023. Corrigendum to "Bubbles and Constraints on Debt Accumulation" [J. Econ. Theory 57 (1992) 245–256]. *Journal of Economic Theory* 208: 105608. Link

Miao, Jianjun and Pengfei Wang. 2018. Asset Bubbles and Credit Constraints. *American Economic Review* 108(9): 2590–2628. Link

Miao, Jianjun and Pengfei Wang. 2025. Responses to "Toward Bubble Clarity: A Comment on Miao and Wang". *Econ Journal Watch* 22(2): 199–210. Link

Pham, Ngoc-Sang and Alexis Akira Toda. 2026. Comment on 'Asset Bubbles and Overlapping Generations'. *Econometrica* 94(3): 1027–1044. Link

Stachurski, John and Alexis Akira Toda. 2019. An Impossibility Theorem for Wealth in Heterogeneous-agent Models with Limited Heterogeneity. *Journal of Economic Theory* 182: 1–24. Link

Stachurski, John and Alexis Akira Toda. 2020. Corrigendum to "An Impossibility Theorem for Wealth in Heterogeneous-agent Models with Limited Heterogeneity" [Journal of Economic Theory 182 (2019) 1–24]. *Journal of Economic Theory* 188: 105066. Link

Tirole, Jean. 1985. Asset Bubbles and Overlapping Generations. *Econometrica* 53(6): 1499–1528. Link

About the Author

Alexis Akira Toda is Professor of Economics at Emory University. His research interests include macro-finance, asset price bubbles, power law, mathematical economics, and computational economics. His papers have appeared in *Econ Journal Watch*, *Econometrica*, *Journal of Applied Probability*, *Journal of Econometrics*, *Journal of Economic Theory*, *Journal of Financial Economics*, *Journal of Monetary Economics*, *Journal of Political Economy*, *Physical Review E*, *Proceedings of the National Academy of Sciences*, *Quantitative Economics*, *Review of Financial Studies*, *SIAM Journal on Numerical Analysis*, *Theoretical Economics*, among others, and he is a co-editor of *Journal of Mathematical Economics*. For more information, visit his website and blog (link). His email address is alexis.akira.toda@emory.edu.

Go to archive of Watchpad section
Go to September 2026 issue