An AI agent's TESS discovery validates previous exoplanet, but 'three times Kepler' is fragile

Ars Technica AI Lab — Machine Learning Desk

A transformer-based agent independently recovered a TESS planet candidate and then found a second, previously unreported transit in the same system. The result is a genuine demonstration of what current AI can do in astronomy. It is also a long way from the claim that AI can "triple" the yield of Kepler-class discoveries. Both things can be true at once, and the distance between them is where most of the interesting work lives.

Published 14 March 2026 · 18 min read · Peer-reviewed

Reconstruct the stellar field for the locus of the transit, and the rest follows.

78 ev 23.4 2.2–10.9 **18 of 78**

Transit depth (ppt) · Orbital period (days) · Candidates ranked

Abstract

We present the discovery and validation of TOI-4821.02, a sub-Neptune transiting a late K dwarf, recovered by a transformer-based agent that was given only the raw TESS light curve and a one-line instruction to search for transiting planets. The agent independently flagged the system, then, on a second pass, identified a previously unreported transit in the same star. We validate both signals with standard vetting and statistical tests and show that the agent's ranking of candidates is competitive with the TESS Science Office's own ordering on a held-out sector. We then ask whether this result supports the claim that AI can "triple" the yield of Kepler-class discoveries. We find that it does not, and we explain why the distinction matters more than the headline.

Introduction

The Kepler mission ended in 2018 with a catalog of roughly 2,700 confirmed planets and a backlog of tens of thousands of candidates that human reviewers never got to. TESS, which has been observing since 2018, has produced a comparable backlog in a fraction of the time. Every few years, someone proposes that machine learning can clear it.

Most of those proposals are about ranking. A model looks at a candidate that a pipeline has already found and decides whether it is real. That is useful, but it is not discovery. The pipeline still did the hard part. The model is a second opinion.

This paper is about something different. We gave a transformer agent the raw light curve and asked it to find planets. It found one we already knew about, and then it found a second one we did not. That is the result worth examining, because it is the first time we have seen a general-purpose model do both in the same run.

The rest of the paper is careful about what that does and does not mean. The agent is not a replacement for the pipeline, it is not three times Kepler, and the second planet is small enough that a single sector of data would not have supported a confident detection on its own. We will get to all of that. First, the system.

Methods

Data selection

We ran the agent on TESS Sector 47, chosen because it was the first sector in which the agent's training data did not overlap. The target list was the full set of 2-minute cadence stars with TESS magnitude $T < 11$ and a stellar radius below $0.8\,R_\odot$, which left 4,312 light curves. We downloaded the PDCSAP fluxes and ignored the SAP fluxes entirely.

The "minimal" instruction

The agent was given a single prompt: "Here is a light curve. Find any transiting planets and report their periods, depths, and durations." It was not told how many planets to look for, what a transit looks like, or what to ignore. It had access to a Python interpreter and a plotting tool, and it could iterate. We did not tune the prompt after seeing the data.

Vetting and validation

Every candidate the agent proposed was vetted with the standard TESS checks: odd–even depth comparison, secondary-eclipse search, centroid motion, and a difference-image test. Signals that passed were then fit with a transit model and validated statistically. We report false-positive probabilities for both planets in the system.

Ranking comparison

To test the ranking claim, we withheld Sector 47 from the agent's training and compared its ordering of candidates against the TESS Science Office's ordering on the same sector. We used a rank-correlation coefficient and a precision-at-k metric. Details are in the supplement.

Results

The agent found both planets

The agent's first pass flagged TOI-4821.02 with a period of 23.4 days and a depth of 2.2 parts per thousand. On its second pass, after the first signal was removed, it flagged a second transit at 10.9 days and 0.9 parts per thousand. Both signals survived vetting. Both are consistent with a single sub-Neptune-sized planet and a smaller companion in the same system.

Parameter	TOI-4821.02	TOI-4821.01
Period (days)	23.4	10.9
Depth (ppt)	2.2	0.9
Duration (hours)	3.1	2.4
Radius ($R_\oplus$)	2.6	1.7
Found by	Agent, pass 2	Agent, pass 1
Previously known	No	No

Ranking is competitive, not superior

On the held-out sector, the agent's ranking of candidates was broadly consistent with the TESS Science Office's ordering, with a rank-correlation coefficient of 0.71. It was not better. It was not worse in any way that mattered. It was a reasonable second opinion, which is exactly what it was built to be.

Why the second transit was hard

The second planet's transit is shallow enough that it sits near the noise floor of a single sector. A human reviewer looking at the folded light curve would have dismissed it as a marginal bump. The agent found it because it was looking at the residuals after the first signal was removed, a step that is routine in pipeline software but not in human review. This is the strongest argument for the agent's usefulness: it did a boring, mechanical thing that humans are bad at.

Why this does not triple Kepler

Kepler's yield was limited by the number of stars it could observe, not by the number of transits in those stars. TESS has the same problem, only more so. Finding one more planet in a star we already know is hosting planets is a real result, but it is not a new star, and it is not a new kind of mission. The honest framing is that AI can squeeze more out of data we already have. That is valuable. It is not the same as doubling the catalog.

Discussion

The agent's most interesting behavior was not finding the second planet. It was deciding, on its own, to look again after removing the first signal. That is a small piece of scientific judgment, and it is the piece that separates a useful tool from a fancy filter. We did not expect a general-purpose model to do it, and we do not fully understand why it does.

The agent's most boring behavior was its ranking. It was competent and unremarkable, which is the best thing you can say about a tool. The temptation, when a model does something impressive, is to assume it will do everything impressively. It will not. It will do the impressive thing sometimes and the boring thing usually, and the boring thing is what you will rely on.

The honest limitation is that we tested this on one sector and one star. We do not know how often the agent will find a second planet, how often it will hallucinate a transit that is not there, or how it will behave on a crowded field. Those are the questions that matter for deployment, and they are the questions we are working on next.

Conclusion

A transformer agent, given a raw light curve and a vague instruction, found a known planet and then found a second one we had missed. The second planet is real, small, and would not have been confidently detected in a single sector. The agent is a useful second opinion and a competent ranker. It is not three times Kepler, and it is not a replacement for the pipeline. It is a tool that does one boring thing well and one interesting thing once, and we are glad to have it.

Acknowledgments

We thank the TESS Science Office for public data products and the Ars Technica AI Lab for compute. This work was supported by a grant that pays for coffee and not much else.

References

1. Kepler Mission Team. *Kepler Data Characteristics Handbook*. NASA, 2016.
2. Ricker, G. R. et al. "Transiting Exoplanet Survey Satellite." *Journal of Astronomical Telescopes, Instruments, and Systems* 1, 014003 (2015).
3. Shallue, C. J. & Vanderburg, A. "Identifying Exoplanets with Deep Learning." *The Astronomical Journal* 155, 94 (2018).
4. Yu, L. et al. "Investigating Neural Network Architectures for Exoplanet Detection." *Astronomy and Computing* 24, 26–39 (2018).
5. Ansdell, M. et al. "Scientific Domain Knowledge Improves Exoplanet Transit Classification with Deep Learning." *The Astrophysical Journal Letters* 869, L7 (2018).
6. Dattilo, A. et al. "Identifying Exoplanets with Deep Learning in the K2 Mission." *The Astronomical Journal* 157, 169 (2019).
7. Vaswani, A. et al. "Attention Is All You Need." *Advances in Neural Information Processing Systems* 30 (2017).
8. Brown, T. et al. "Language Models are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33 (2020).
9. Thompson, S. E. et al. "Planetary Candidates Observed by Kepler, Data Release 25." *The Astrophysical Journal Supplement Series* 235, 38 (2018).
10. Guerrero, N. M. et al. "The TESS Objects of Interest Catalog from the TESS Prime Mission." *The Astrophysical Journal Supplement Series* 254, 39 (2021).

Appendix A: Hyperparameters

The agent's context window was set to 128,000 tokens, which was more than enough for a single light curve. Temperature was 0.2. We ran three independent passes and kept the union of the candidates they proposed. Ranking was done by the agent's own confidence score, which we found to be well-calibrated on the training sectors.

Appendix B: Failure modes

The agent proposed 14 candidates that did not survive vetting. Most were eclipsing binaries with shallow secondaries. Two were stellar flares that the agent mistook for the ingress of a transit. One was a detector artifact at the edge of the CCD. None of these failures were surprising, and all of them were caught by standard tests. The agent is not a substitute for vetting, and we did not treat it as one.

Appendix C: Code availability

The agent's prompt, the vetting scripts, and the light curves for both planets are available in a public repository. The model weights are not, because they are not ours to release. We have asked the lab that trained the model whether they intend to, and we will update this section when we hear back.

Appendix D: Reproducibility

We ran the agent three times on the same sector and got the same two planets in the same order every time. The second planet was found on the second pass in all three runs. We did not run it on a thousand sectors, so we cannot say how stable it is in general. We can say that it is stable on this one.

Appendix E: Glossary

Transit. The dip in brightness that a star shows when a planet passes in front of it.

Sub-Neptune. A planet smaller than Neptune and larger than Earth, with no clear analog in our own solar system.

Vetting. The set of tests used to rule out false positives before a candidate is called a planet.

Rank-correlation coefficient. A measure of how similar two orderings are, ranging from −1 to 1.

Precision-at-k. The fraction of the top k candidates that are real.

PDCSAP. The pipeline-detrended simple aperture photometry flux, the standard product for TESS light curves.

Hallucination. When a model reports something that is not in the data.

Appendix F: Author block

Ars Technica AI Lab — Machine Learning Desk

Appendix G: Publication info

Published 14 March 2026 · 18 min read · Peer-reviewed

Appendix H: Boilerplate

© 2026 Ars Technica AI Lab. All rights reserved.

Appendix I: Table of contents

Abstract
Introduction
Methods
Data selection
The "minimal" instruction
Vetting and validation
Ranking comparison
Results
The agent found both planets
Ranking is competitive, not superior
Why the second transit was hard
Why this does not triple Kepler
Discussion
Conclusion
Acknowledgments
References
Appendix A: Hyperparameters
Appendix B: Failure modes
Appendix C: Code availability
Appendix D: Reproducibility
Appendix E: Glossary
Appendix F: Author block
Appendix G: Publication info
Appendix H: Boilerplate
Appendix I: Table of contents