

Peer Review

Review of: "PhysAgent: A Multi-Agent Approach to the Automated Discovery of Physical Laws"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper presents PhysAgent, a three-role system built on large language models. A Mentor Agent asks guiding questions, a Student Agent writes and runs code (Python, ASE, Quantum ESPRESSO), and a Leader Agent routes tasks between the two; a human user supplies data and feedback. The authors say the system "autonomously" carries a physics project from hypothesis to result, and they support this with four case studies: Kepler's three laws recovered from simulated trajectories of five planets (Figure 2), Newton's second law recovered from simulated position-time data of a cart (Figure 3), a DFT electronic-structure workflow for GaAs (Figure 4, Appendix step list, Listings 1 and 2), and a thin-film simulation of rain running across a train window (Figure 5, Eqs. 1–6, Table 3). Table 1 and Table 2 compare PhysAgent's capabilities with AutoGPT, LangChain, and MetaGPT. The Methods section reproduces the role prompts, partly truncated.

The strongest claims sit in the Abstract, the last paragraph of the Introduction, and the first paragraph of the Discussion: that the laws were derived "without prior knowledge", that the system "ensures reproducibility", and that it "establishes a trustworthy, scalable paradigm". I think the engineering idea is sensible, and the paper is honest in places where many agent papers are not. The Discussion names concrete limitations (reliance on simulated data, human intervention for figures and writing, trouble with open-ended problems). The authors also printed the prompts and a real DFT parameter set, which lets a reader check what the agents were actually told. That transparency is what makes most of my critique below possible, and I want to credit it.

2. Whether the conclusion holds on the data

There is no quantitative evaluation anywhere in the paper. I read the whole Results section, Methods, Appendix, and the three tables. The evidence consists of one successful run per task, shown as figures. I could not find the number of attempts per task, a success rate, the number of failed runs, the number of human interventions, wall-clock time, token cost, or which of the five listed LLMs (GPT-4, Kimi, Doubao, Llama, DeepSeek, Figure 1 caption) drove which case study. No baseline is reported either: not a single-agent run with the same model, not a run without the Mentor, not a human analyst. Without at least one such comparison, the statement in the Results subsection on architecture that PhysAgent "demonstrates significant advantages" over "traditional single-agent models and general multi-agent systems" has no measurement behind it. Tables 1 and 2 are feature checklists filled in by the authors, not experiments.

"Without prior knowledge" is contradicted by the paper's own prompts. This is the central claim of the Kepler and Newton case studies, and the Methods section undermines it directly. The Mentor Agent prompt says: "Rather than directly revealing target laws (such as Newton's second law), you lead students to discover them". The Student Agent prompt tells the model its goal is "deriving an unknown physical law (e.g., Newton's second law)" and names the data to collect: "force, acceleration, mass". The Leader Agent prompt repeats "deriving physical laws (e.g., Newton's Second Law)" twice. The appendix list of Mentor comments includes item 14, "Avoid using formulas that presuppose $F=ma$ ", and item 16 recommends "Galileo's inclined plane and Atwood machine". So the target law is written into the system prompts, and the underlying models were trained on text that contains it. The Kepler section is candid about this in one sentence: the student agent formulated hypotheses "Based on prior knowledge". The Mentor's quoted question, "Circular or elliptical?", also hands over the answer space. What the case studies show is that an LLM pipeline can confirm a known law on clean synthetic data when steered toward it. That is a reasonable result. It is not discovery without prior knowledge, and the Abstract should not say so.

The Newton experiment cannot support the linear relations it reports. The text under Figure 3 lists six (F, m) pairs: (0.1 N, 2.0 kg), (0.2 N, 0.6 kg), (0.3 N, 0.2 kg), (0.4 N, 0.8 kg), (0.4 N, 2.0 kg), and (0.5 N, 0.6 kg). I grouped them. At fixed mass, $m = 0.6$ kg has two forces (0.2 and 0.5 N), $m = 2.0$ kg has two (0.1 and 0.4 N), and $m = 0.2$ kg and $m = 0.8$ kg have one each. At fixed force, only $F = 0.4$ N has two masses (0.8 and 2.0 kg); every other force has a single point. Figure 3(c) claims "a clear linear relationship between acceleration and force" within each mass group, and Figure 3(d) claims "a strong linear dependence" on $1/m$ at fixed force. With at most two points per group, any two points define a line; linearity is not

testable, and the $a \propto 1/m$ claim rests on one pair of runs. The Mentor's own advice in the appendix (item 11: "confidence intervals and residual tests"; item 12: "Eliminate alternative mathematical relationships") was, as far as I can see, not followed in the reported result, since no slope, intercept, R^2 , residual, or interval is given. (As rendered on Qeios, the mass and force lists in the paragraphs describing Figure 3(c) and 3(d) are partly lost, so I used the six pairs listed under Figure 3(b).) The data are also simulated by the human user, presumably from $F = ma$ plus "small amounts of noise," and the noise level is not stated. The recovery is therefore of the generating equation, which should be said plainly. One more point: the cart is on an "inclined plane," so gravity has a component along the track, and the paper never says whether F is the applied force or the net force.

The Kepler study is described only qualitatively. Figure 2(c) says the orbits "are elliptical with the Sun at one focus," but no eccentricity, semi-major axis, focus position, or goodness of fit is reported, and I could not find how "elliptical" was decided against "circular." Figure 2(d) says areal velocity is "approximately constant" with no stated tolerance. Figure 2(e) fits T^2 against a^3 on five points (as rendered on Qeios, the formula in that caption is not legible). How the trajectories were generated (integrator, time step, units, whether planet-planet interactions were included, noise) is not given, so I cannot judge how hard the inference was.

The GaAs case is a workflow execution, and the headline number is missing. The Abstract and Introduction promise "GaAs band gaps," yet I could not find a band-gap value anywhere in the text. Figure 4(f) concludes only that GaAs is "a direct band gap semiconductor." The exchange-correlation functional and the pseudopotentials are not named (Appendix step 4 says only "Pseudopotential files for Ga and As in UPF format"), so a reader cannot tell whether the unreported gap would be expected to agree with experiment. The Appendix gives $ecutwfc$ 40 Ry, $ecutrho$ 320 Ry, $degauss$ 0.01, a k -grid "of 8" for SCF and 12 for NSCF, and a lattice constant of 5.65 Å; no convergence test with respect to cutoff or k -points is shown, only SCF energy against iteration in Figure 4(c). There is also an inconsistency in how the structure was built. Figure 4(a)–(b) says the user uploaded a CIF file that ASE parsed, whereas Appendix step 2 says the structure was made with ASE's `bulk` function at 5.65 Å. Appendix step 1 advises to "search online for solutions or seek assistance from experienced users" if installation fails, which suggests human help that the case study does not count. The Abstract names VASP as an integrated tool; I found no VASP calculation in the paper.

The rain-on-window simulation has physical and numerical problems. I checked Eqs. 1–6 and Table 3 and found four issues.

- *Gravity sign.* From Eq. 1 (right-hand side $-\nabla p + \mu \nabla^2 u + \rho g$), the lubrication balance is $\mu \partial^2 u / \partial z^2 = \partial p / \partial y - \rho g_y$. Eq. 2 writes $+\rho g_y$. Carried into Eq. 3, with $g_y = -g$ as stated, the mean velocity term becomes $+\rho g h^2 / (3\mu)$, i.e. upward along y , which the text defines as vertical height. Eq. 6 then uses $\nabla(\sigma \nabla^2 h - \rho g y)$, which does give downward flow, so Eq. 6 is not consistent with Eqs. 2–4 as printed.
- *Dimensional error in the damping term.* The text replaces $\nabla(\sigma h^3 / (3\mu) \nabla^2 h)$ by $\alpha \sigma h^3 / \mu \nabla^2 h$ "(where α is a dimensionless damping coefficient)". The first expression has units of m/s ($\sigma h^3 / \mu$ is m^4/s , times m^{-3}). The replacement has $\sigma h^3 / \mu$ times $\nabla^2 h$, i.e. m^4/s times $m^{-1} = m^3/s$. For the two to match, α must carry units of m^{-2} . It cannot be dimensionless. The replacement also turns fourth-order capillary smoothing into second-order diffusion, which is a different physical model, not an approximation of the same one. Its value is not in Table 3.
- *Air shear magnitude.* Table 3 gives $\tau_{\text{air}} = 36600$ Pa with $\mu = 1 \times 10^{-3}$ Pas. Using the shear term of Eq. 3, $h \tau_{\text{air}} / (2\mu)$, at the smallest film thickness that is allowed to move ($h_{\text{min_move}} = 1 \times 10^{-5}$ m, Numerical Methods section): $(1 \times 10^{-5} \times 36600) / (2 \times 1 \times 10^{-3}) = 183$ m/s. The film Reynolds number at that point is $\rho u h / \mu = 1000 \times 183 \times 1 \times 10^{-5} / 1 \times 10^{-3} = 1830$, which contradicts the low-Reynolds, negligible-inertia assumption stated before Eq. 2. With $dx = 0.0025$ m and $dt = 0.0002$ s, the advective Courant number is $183 \times 0.0002 / 0.0025 = 14.6$. The time-integration scheme is not stated, so I cannot say whether that is unstable, but it needs explaining.
- *Thin-film validity.* The dripping threshold in Table 3 is 8×10^{-2} m, an 8 cm film, which is not thin by any reading of $h \ll L$. The evaporation rate of 1×10^{-4} m/s removes a 1×10^{-5} m film in 0.1 s; I would like to see that value justified.

Beyond the equations, the branching and meandering that make Figure 5(e)–(f) resemble the photographs in 5(c)–(d) are inserted by rule: "random fluctuations" on the main path, a branch probability of 0.85 (Table 3), random secondary and tertiary branches, random drops. The tree-like patterns are therefore inputs, not outputs of the thin-film equation, and "reproduces the flow trajectories" is supported only by visual resemblance. No quantitative comparison with the observations (angle, speed, branch spacing) is offered. The section heading promises to "simulate and interpret" the phenomenon; I did not find an interpretation.

The Leader Agent's contribution is asserted, not shown. The Abstract says it "dynamically optimizes task scheduling and resource allocation". The prompt in Methods describes routing tasks and alternating between two members; none of the case studies reports a scheduling decision, a resource measure, or a

comparison with fixed alternation. Similarly, the Introduction lists "safety mechanisms that enforce physical consistency, such as energy conservation checks" as a requirement; I found no such mechanism in PhysAgent, only a Mentor suggestion (appendix item 8).

Overall, the claims in the Abstract and Discussion ("robust capabilities", "trustworthy, scalable paradigm", "ensures reproducibility", "validate PhysAgent") go well beyond four single demonstrations on synthetic or tutorial-level tasks.

3. Reproducibility

Some pieces are present, and they help: the role prompts (partly), the full DFT step list with cutoffs and k-grids, a Quantum ESPRESSO path that reveals version 7.4.1 (qe-7.4.1/pseudo, Appendix step 3), and a complete parameter table for the thin-film run (Table 3). The rest is missing.

Models and decoding. As is usual for LLM-agent papers, the model identity is the largest gap. Figure 1 lists five model families with no versions or snapshot dates, and no case study says which one was used. Temperature, top-p, maximum tokens, context length, and seeds are not reported. Because outputs of these systems vary from run to run, the absence of repeated runs means the reader cannot tell whether each figure is typical or the best of many.

Prompts. The Methods prompts end in "....." (Student item 7, Leader items 6–10), and the text says they "are not yet complete". The task-level prompts the human gave, the Mentor's full dialogue, and the Student's intermediate code are not provided. The Methods label the Mentor prompt as "role requirement prompts related to the Leader Agent," which looks like a copy error.

Code and data. The Methods call the GitHub link "the future open-source GitHub repository," while the Code availability statement says "all code, datasets, and related materials from this study are publicly accessible" there. Those two sentences cannot both be true as of the same version. I did not check the repository and am reviewing the Qeios text only. As rendered on Qeios, Listing 1 and Listing 2 show captions but no code.

Human involvement. The human generated all data for three of the four cases, gave feedback "at key stages" on figures and writing (Newton section), and may have assisted with installation (Appendix step 1). None of this is logged or counted, so the degree of autonomy cannot be reproduced or assessed.

Physics inputs. For DFT: functional, pseudopotential files, and the k-grid shift are missing. For the thin film: α , the time-integration scheme, initial condition, and the random seed for branching and drops are

missing. For Kepler and Newton: the generating equations, integrator, noise model, and the data files.

4. What specifically to fix

1. **Abstract, Introduction final paragraph, Kepler and Newton sections.** Remove "without prior knowledge" and "purely from observational data". The prompts in Methods name Newton's second law and the variables force, acceleration, and mass, and the Kepler section says hypotheses came "Based on prior knowledge". Recast the result as guided confirmation of known laws on synthetic data.
2. **Add a quantitative evaluation.** For each of the four tasks, run N independent attempts (say 10) per underlying model and report success rate, failure modes, number of human interventions, tokens, and time. Include at least a single-agent baseline and a no-Mentor ablation using the same model. This is what would turn the Table 1 checklist into evidence.
3. **State the model per experiment,** with exact version or snapshot, temperature, top-p, max tokens, and seed, in a table next to Table 3.
4. **Figure 3 and the (F, m) list under Figure 3(b).** Redesign the Newton dataset as a full grid (several forces at each of several masses, for example 4×4), report fitted slopes with confidence intervals, intercepts, R^2 , and residual plots, and compare $a \propto F$ against at least one alternative form. With the current six pairs, no group has more than two points. Say whether F is applied or net force on the incline, and give the noise level.
5. **Figure 2:** report fitted eccentricities and semi-major axes, the tolerance used to call areal velocity constant, and the fitted slope of T^2 against a^3 with its uncertainty. Describe how the trajectories were generated.
6. **Figure 4 and Appendix.** Either report the computed GaAs band gap with functional and pseudopotentials named, compared against a reference value with the expected functional error discussed, or remove "band gaps" from the Abstract and Introduction. Reconcile the CIF upload in Figure 4(a)–(b) with the `bulk(..., 5.65 Å)` construction in Appendix step 2. Add cutoff and k-point convergence checks. Remove VASP from the Abstract unless a VASP run is shown.
7. **Eqs. 2–4:** fix the sign of the gravity term so that Eqs. 2–4 and Eq. 6 agree (downward flow for $g_y = -g$). Clarify the sign convention for τ_{air} , since the text calls it negative-x and Eq. 3 also subtracts it.
8. **Paragraph after Eq. 6:** α has units of m^{-2} , not dimensionless. Give its value, and acknowledge that the substitution changes the model from capillary to diffusive smoothing.

9. **Table 3, $\tau_{\text{air}} = 36600 \text{ Pa}$.** Justify the value or correct it; at $h = 1 \times 10^{-5} \text{ m}$ it gives a film velocity of 183 m/s and a Reynolds number of 1830, and a Courant number of 14.6 at the stated dx and dt . State the time-stepping scheme. Also justify the $8 \times 10^{-2} \text{ m}$ dripping threshold under a thin-film assumption.
10. Figure 5(e)–(f): say explicitly that branching, meandering, and drops are imposed stochastically, and add at least one measured comparison with the observations in 5(c)–(d), for example, streak angle or branch spacing. If the model has no free parameter tied to train speed, say so.
11. **Leader Agent:** show one logged scheduling decision and compare against simple alternation, or reduce the claim in the Abstract to "routes tasks between agents".
12. **Methods:** fix the "related to the Leader Agent" label above the Mentor prompt, publish the complete prompts in place of the "....." truncations, and make the Code availability statement consistent with "the future open-source GitHub repository". Include the code of Listings 1 and 2 in a form that renders on Qeios.
13. **Discussion, first paragraph:** replace "robust capabilities", "trustworthy, scalable paradigm", and "validate" with wording that matches four single-run demonstrations.

5. Recommendation

Major revisions. The system is a reasonable piece of engineering, and the case studies are useful as demonstrations; the printed prompts and parameter tables are a real strength. But the paper's main claim, autonomous discovery without prior knowledge, is contradicted by its own prompts, there is no repeated-run evaluation or baseline, the Newton dataset has at most two points per group, the GaAs band gap is promised and not reported, and the thin-film model contains a sign error, a dimensional error, and an air-shear value that breaks the model's own assumptions. All of this is fixable. It needs new experiments and a narrower framing, not only editing.

Declarations

Potential competing interests: No potential competing interests to declare.