STARGAZER: A Scalable Model-Fitting Benchmark Environment for AI Agents under Astrophysical Constraints

Xinge Liu [*1] **Terry Jingchen Zhang** [*2] **Bernhard Schölkopf** [3 4] **Zhijing Jin** [1 2 3] **Kristen Menou** [1]

Abstract

The rise of autonomous AI agents suggests that dynamic benchmark environments with built-in feedback on scientifically grounded tasks are needed to evaluate the capabilities of these agents in research work. We introduce STARGAZER, a scalable environment for evaluating AI agents on dynamic, iterative physics-grounded model-fitting tasks using inference on radial-velocity (RV) time series data. STARGAZER comprises 120 tasks across three difficulty tiers, including 20 real archival cases, covering diverse scenarios ranging from high-SNR single-planet systems to complex multi-planetary configurations requiring involved low-SNR analysis. Our evaluation of eight frontier agents reveals a gap between numerical optimization and adherence to physical constraints: although agents often achieve a good statistical fit, they frequently fail to recover correct physical system parameters, a limitation that persists even when agents are equipped with vanilla skills. Furthermore, increasing test-time compute yields only marginal gains, with excessive token usage often reflecting recursive failure loops rather than meaningful exploration. STARGAZER presents an opportunity to train, evaluate, scaffold, and scale strategies on a model-fitting problem of practical research relevance today. Our methodology to design a simulation-driven environment for AI agents presumably generalizes to many other model-fitting problems across scientific domains. The benchmark implementation and accompanying project resources are available at `https://github.com/AIPS-UofT/Stargazer` and `https://aips-uoft.github.io/Stargazer`.

[*]Equal contribution [1]University of Toronto [2]Vector Institute [3]Max Planck Institute for Intelligent Systems, Tübingen, Germany [4]ELLIS Institute Tübingen. Correspondence to: Kristen Menou <kristen.menou@utoronto.ca>. Email: Xinge Liu <xinge.liu@mail.utoronto.ca>.
Preprint. June 8, 2026.

1. Introduction

> *"If planets are colossally abundant, as it seems they are, then perhaps life is too."*
>
> **Michel Mayor**
> Nobel Laureate (2019), discoverer of the first exoplanet

Mastering the laws of physics has long been considered a defining challenge for artificial intelligence, as the discipline of physics demands tight integration of experimental observation with theoretical derivation (Wang et al., 2023; Krenn et al., 2022). While frontier models have made rapid progress on question answering (QA) benchmarks such as HLE (Center for AI Safety et al., 2026) and GPQA (Rein et al., 2024), AI for scientific discovery increasingly calls for agentic, multi-step workflows that involve tool-calls, simulations, and iteratively learn from feedback over repeated attempts (Shen et al., 2026; Chen et al., 2024; Lupidi et al., 2026). However, existing agentic benchmarks largely focus on simplified environments such as equation discovery (Zheng et al., 2026; Koblischke et al., 2025) that do not fully capture the complexity of real-world research. We aim to go one step further in emulating high-fidelity scientific workflow of frontline researchers in the field of astrophysics.

Exoplanet discovery is tied to one of the most critical existential questions for humanities: are we alone in the universe? A concrete step toward answering it is to find Earth-like planets with potentially habitable environments (Perryman, 2018). Since the first exoplanet discovery in 1995 (Mayor & Queloz, 1995), radial-velocity (RV) spectroscopy has provided a robust dynamical way to indirectly detect planets even when they do not pass in front of their host star. With more than six thousand confirmed exoplanets (Winn & Fabrycky, 2015), RV methods remains a cornerstone for characterizing planetary systems.

We introduce STARGAZER, a high-fidelity testbed that evaluates frontier agents on an autonomous scientific workflow of exoplanet discovery using radial velocity (RV) methods. RV analysis is an ideal platform for evaluating scientific agents for three reasons. First, it demands a structured, multi-step workflow, including periodogram analysis, iterative Keplerian fitting, model selection, and submission, that

cannot be short-circuited by retrieval or pattern matching. Second, success is objectively verifiable: a proposed planetary configuration either matches the ground truth or it does not. Third, task complexity scales with physics grounding, allowing fine-grained difficulty control without artificial contrivance.

STARGAZER contains an infinitely scalable data synthesis pipeline with 120 sample tasks spanning three difficulty levels, including 20 tasks drawn from real archival stellar spectra. We evaluated 8 frontier models and report three findings: (1) statistical fit quality does not imply physical recovery; (2) token usage does not predict performance; and (3) successful agents escalate model complexity while failed agents repeat. Bootstrapped skills help on Easy-tier tasks, but do not reliably transfer to Hard-tier tasks. Critically, these challenges highlight that our simulation-driven feedback framework addresses a fundamental bottleneck in automated inference, one that presumably generalizes to diverse model-fitting problems across scientific domains. Overall, STARGAZER represents a step towards accelerating exoplanet discovery with AI agents and offers insights for future practitioners.

2. Related Work

LLM Benchmarks in Physics and Astronomy. Physics is gaining increasing prevalence in LLM capability evaluation. General scientific-reasoning suites such as SCIBENCH (Wang et al., 2024), OLYMPIADBENCH (He et al., 2024), and GPQA (Rein et al., 2024) feature physics as a core domain, and dedicated physics reasoning benchmarks (Xu et al., 2025; Zhang et al., 2025; Xiang et al., 2025; Zhu et al., 2025; Zhao et al., 2026) have further shown that model performance degrades sharply as problems demand more sophisticated physical theorems and derivations. This trend extends to astronomy, where ASTROMLAB 1 (Ting et al., 2024) evaluates expert-level knowledge retrieval and ASTROMMBENCH (Shi et al., 2025) broadens the scope to multimodal image interpretation. Despite this progress, existing benchmarks uniformly draw from coursework, competitions, or factual recall, leaving open whether models can reason over raw empirical data as a working researcher would.

LLM Agents for Physics Research. Recent work explores LLM-based agents for physics research, spanning both symbolic/theoretical reasoning and data-driven discovery. Brenner et al. (2026) show the neuro-symbolic progress on an open theoretical physics problem. On the empirical side, existing benchmarks in the physical sciences focus on narrow slices of the workflow: ASTROVISBENCH (Joseph et al., 2025) addresses a single astronomy stage; GRAVITY-BENCH (Koblischke et al., 2025) studies gravitational-law discovery from simulation but omits scalable real observations; and ASTROREASON-BENCH (Wang et al., 2026) emphasizes mission-level planning. Complementary efforts emphasize tool use and simulation execution under practical constraints, including SIMULCOST (Cao et al., 2026), PHYSICSMIND (Mak et al., 2026), and SCIAGENT-GYM (Shen et al., 2026). More general agent benchmarks assess components that physics agents also require, for example code generation (Chen et al., 2024), data-driven scientific discovery (Chen et al., 2025; Lupidi et al., 2026), hypothesis search (Majumder et al., 2024), reproducibility (Siegel et al., 2024), and workflow execution (Tian et al., 2024), but do not test end-to-end reasoning from raw measurements to physical interpretation. STARGAZER fills this gap by challenging agents to execute end-to-end exoplanet-discovery workflows using radial-velocity methods, requiring them to process noisy time-series data, select appropriate physical models, and extract orbital parameters at the complexity real-world astrophysical research demands.

3. STARGAZER

Physics-Grounded Environment. STARGAZER simulates the problem of exoplanet discovery and characterization from stellar radial velocity (RV) observations. We synthesize RV time series by modeling the gravitational influence of orbiting planets on their host star. Each planetary system is parameterized by orbital period, eccentricity, argument of periastron, and orbital phase, which together determine the velocity signal induced on the star.

The stellar reflex motion is modeled using Keplerian orbital dynamics and can optionally incorporate full N-body integrations for multi-planet systems when dynamical interactions become significant. The resulting RV signal is sampled at irregular observation times to reflect realistic telescope schedules and observational constraints. Measurement noise and stellar activity are incorporated through Gaussian observational uncertainty and correlated noise processes. Formally, the observed radial velocity signal at time t is modeled as

$$v(t) = \sum_{i=1}^{N_p} v_i(t; \theta_i) + \gamma + \epsilon(t), \tag{1}$$

where N_p denotes the number of planets in the system, $v_i(t; \theta_i)$ represents the Keplerian velocity contribution of planet i with orbital parameters θ_i, γ is the systemic velocity offset of the star, and $\epsilon(t)$ represents measurement and stellar noise. For multi-instrument datasets, a separate offset γ_j is fitted per instrument. The agent is given access only to the observed time series and must infer the underlying planetary configuration.

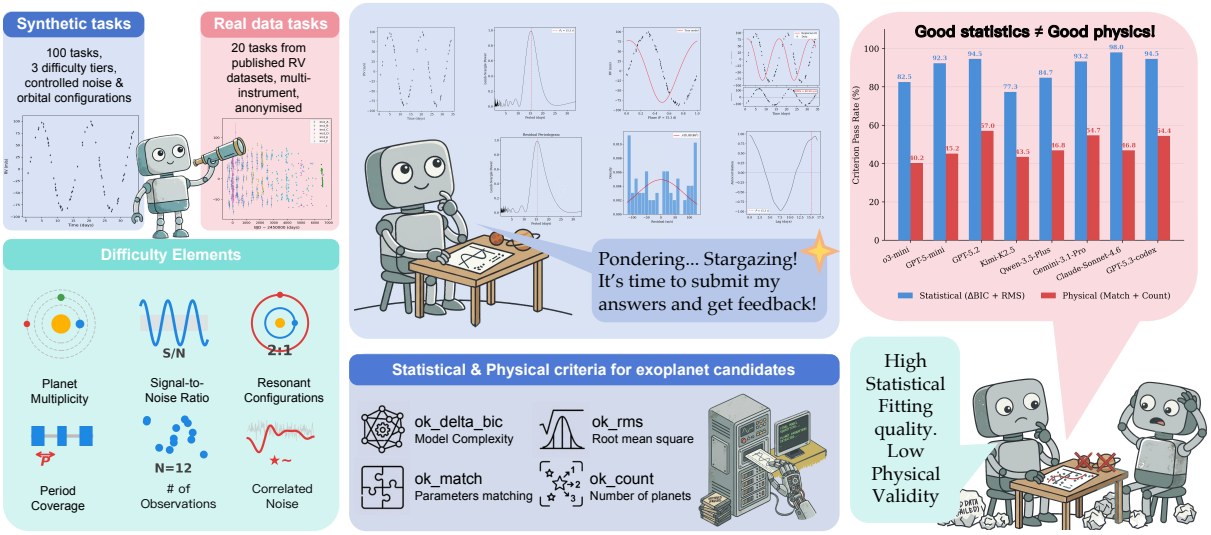

Figure 1. Overview of Stargazer. **Left:** 120 RV tasks (100 synthetic, 20 real), with synthetic difficulty controlled by six physical factors. **Center:** Agents run a periodogram-to-Keplerian workflow and are graded on statistical and physical criteria. **Right:** Models often achieve strong statistical fits but fail to recover correct orbital parameters.

3.1. Task Construction

STARGAZER comprises 100 synthetic and 20 real-data RV tasks, grouped into three difficulty tiers: Easy (20 tasks), Medium (40 tasks), and Hard (40 tasks). As illustrated in Figure 2 (left), each synthetic task is fully determined by a single random seed: the seed controls orbital parameter sampling, observation scheduling, noise injection, and N-body signal generation via REBOUND (Rein & Liu, 2012) (details in Appendix A.1). Because every task is reproducible from its seed alone, the synthetic component is *infinitely scalable*: new held-out suites can be generated on demand, preventing score saturation as models improve. The 20 real-data tasks are constructed from published archival RV datasets; we provide conversion scripts that anonymise and reformat archival data into the STARGAZER interface, making it straightforward to incorporate additional real systems in the future (Appendix A.3).

Difficulty scoring. Each task is assigned an integer difficulty $d \in [1, 10]$ by summing six physics-based components derived from established RV theory and split into three difficulty levels (Cumming, 2004; Anglada-Escudé et al., 2009; Queloz et al., 2001; Haywood et al., 2014): planet multiplicity, SNR, resonant configurations, period coverage, observation count, and correlated noise amplitude (Figure 1, complete rubric in Appendix A.1). Factor weights were set *a priori* from domain knowledge. Tasks that were physically non-identifiable under the realized observation window and noise were filtered out to ensure solvability.

3.2. Agentic Environment

As Figure 2 illustrates, the environment provides the agent with the RV dataset (timestamps, velocities, uncertainties, and host star mass) at the beginning of each episode. The agent operates in a ReAct-style loop with two tools: a **PythonREPL** for executing analysis code (periodograms, Keplerian fitting, residual inspection) and **submit_action** interface for proposing candidate planetary systems. The agent may submit multiple times within an episode; only the best submission counts toward the final score. After each submission, the evaluator returns per-criterion diagnostic signals (pass/fail status and optional hints), enabling the agent to revise its hypothesis, for example by adding a planet if the count is wrong or refining parameters if the match score is low.

Automated grading. The evaluator reconstructs the RV curve from the agent's submitted parameters via forward modeling, then applies four pass/fail criteria detailed in §3.3.

Resource budgets. Each tier has a fixed resource budget calibrated from pilot runs: we observed the token and time consumption of successful trajectories and set limits at approximately 3× the median successful cost to provide ample headroom while preventing runaway episodes. The resulting budgets are 200K tokens / 600 s (Easy), 450K / 900 s (Medium), and 900K / 1500 s (Hard), with 3, 5, and 10 submission attempts respectively. An episode terminates when any limit is reached.

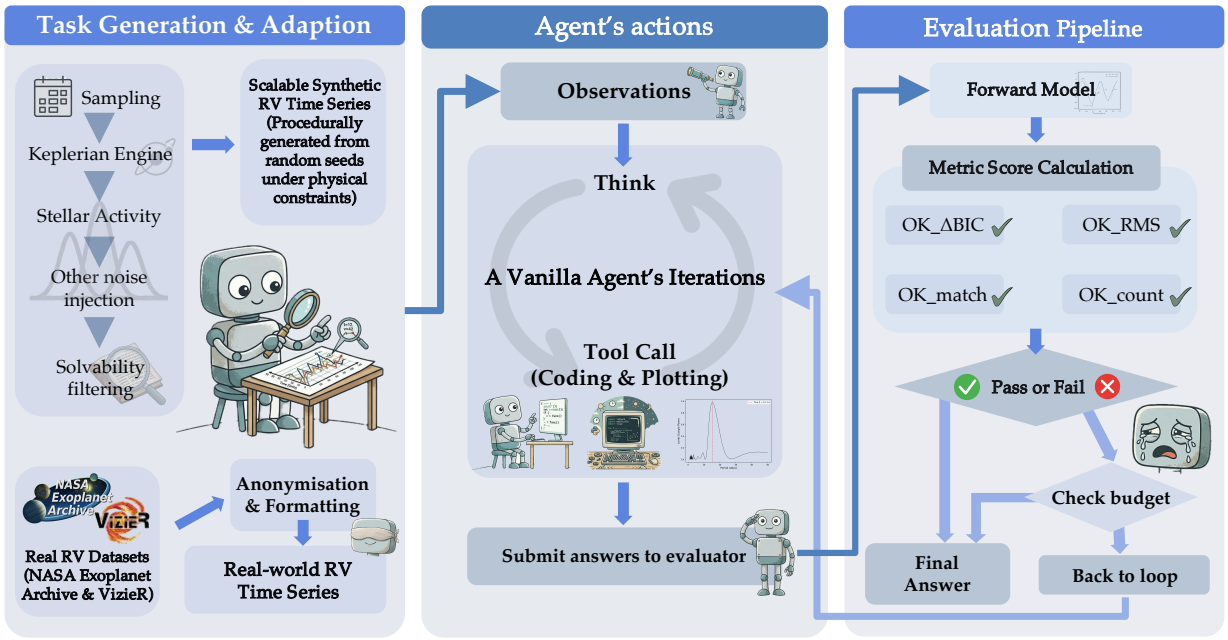

Figure 2. STARGAZER framework. **Left:** Task generation from synthetic physics or extracted from archival RV data. **Center:** Agent iteration loop of analysis, submission, and per-criterion feedback. **Right:** Evaluator forward-models submissions and grades with ΔBIC, RMS, Match, and Count.

3.3. Evaluation Protocol

A task is considered solved only when *all four* criteria below are simultaneously satisfied.

Statistical metric. Given the observed radial velocities $\{y_i\}$, measurement uncertainties $\{\sigma_i\}$, and the agent's predicted model $\{\hat{y}_i\}$, the environment computes:

(1) ok_rms (residual quality): the root-mean-square residual $\mathrm{RMS} = \sqrt{N^{-1}\sum_i (y_i - \hat{y}_i)^2}$ must satisfy $\mathrm{RMS} \leq 1.5\,\tilde{\sigma}$, where $\tilde{\sigma}$ is the median reported measurement uncertainty. This ensures the model fits the data to within a factor of the observational noise floor.

(2) ok_delta_bic (model selection): the per-point $\Delta\mathrm{BIC}/N$ must be positive, where $\Delta\mathrm{BIC} = \mathrm{BIC}_{\mathrm{null}} - \mathrm{BIC}_{\mathrm{model}}$. The null model is a weighted-mean constant. The BIC is computed as $\mathrm{BIC} = -2\ln\mathcal{L} + k\ln N$, with $k = 5n_{\mathrm{pl}} + n_{\mathrm{inst}}$ free parameters (five Keplerian elements per planet plus one systemic velocity per instrument). A positive $\Delta\mathrm{BIC}/N$ confirms that the submitted model is statistically preferred over a flat line, penalising over-parameterised solutions.

Physical metric. Submitted planets are matched to truth planets via the Hungarian algorithm (Kuhn, 1955; Budavári & Basu, 2016; Hopkins et al., 2015) on a pairwise distance matrix:

$$d_{ij} = 4.0\,\frac{\mathrm{RMS}(\mathrm{RV}_i - \mathrm{RV}_j)}{K_i} + 1.0\left|\ln\frac{P_j}{P_i}\right| + 0.5\left|\ln\frac{K_j}{K_i}\right| + 0.5\,|e_j - e_i|, \tag{2}$$

where the RV-curve term compares single-planet Keplerian signals with an optimal offset removed. Its dominant weight (w=4.0) naturally absorbs parameter degeneracies (e.g., the ω/M_0 trade-off at low e). Pairs with $d_{ij} > 5$ are rejected. The match score is

$$S_{\mathrm{match}} = \frac{1}{|\mathcal{M}|}\sum_{(i,j)\in\mathcal{M}} e^{-d_{ij}} - 0.25\left|n_{\mathrm{truth}} - n_{\mathrm{guess}}\right|. \tag{3}$$

(3) ok_match: $S_{\mathrm{match}} \geq 0.8$. The threshold $S_{\mathrm{match}} \geq 0.8$ corresponds to a mean pairwise distance $d \leq 0.22$, requiring that submitted and ground-truth Keplerian signals closely overlap — a criterion more permissive than the parameter precisions typically reported in RV discovery papers (e.g., $\sigma_P/P < 1\%$, $\Delta e < 0.05$ (Lovis et al., 2006a)). The match-score distribution is strongly bimodal (Figure 3, left), with only 14% of submissions in the $[0.72, 0.88]$ boundary region; sweeping the threshold by $\pm 10\%$ shifts pass rates by at most 5.0 pp and preserves all model rankings (Figure 3, right).

(4) ok_count: $n_{\mathrm{guess}} = n_{\mathrm{truth}}$.

The first two criteria verify that the submitted model fits the data; the latter two verify that it recovers the correct physical system. This conjunction gate prevents trivial solutions that achieve low residuals without identifying the right planets. All thresholds are fixed across tasks and models; only the RMS threshold adapts implicitly through $\tilde{\sigma}$, which varies with data quality. The match score is computed as the

mean over successfully paired planets; unmatched planets are handled by the separate count-match criterion rather than included in the score itself. This conjunction gate captures orthogonal failure modes, as a model may pass **ok_match** while failing **ok_count** or vice versa, providing finer discrimination between models than either criterion alone.

4. Results and Discussion

Baselines without LLMs. We include two baselines using deterministic programs without LLMs. The **Classical Pipeline** chains Lomb-Scargle periodogram search, weighted circular-orbit initialization, multi-start Keplerian fitting (`scipy.optimize.least_squares`), and greedy BIC-gated planet addition (ΔBIC $>$ 10). The **Nested Sampling** baseline uses Bayesian model comparison via nested sampling to select the number of planets, then fits orbital parameters under the best model. Both achieve 95.0% on Easy (stronger than the best LLM agents) but degrade on harder tiers (Classical: 35.0% Medium, 5.0% Hard; Nested Sampling: 32.5% Medium, 0.0% Hard). They shared inability to reliably detect more than one planet (average predicted count $\approx$1.1 across all tiers). These baselines demonstrate LLM agents have not yet outperformed traditional methods on simple single-planet tasks.

Performance drops sharply as difficulty increases. As Table 1 shows, top performers exceed 70% on Easy tasks, but drops to 17–35% on Medium and collapses on Hard, where no model exceeds 6%. This degradation is consistent across all models, which confirms the construct validity of our physics-grounded stratification. GPT-5.3-codex achieves the best Easy-tier pass rate (80.0%) and the highest natural completion rate across all tiers, indicating that it reaches termination conditions within budget more often than other models. In contrast, o3-mini completes 76.7% of Easy tasks but has a 0% pass rate on Hard, where it often stop early without producing correct results. We also found that the dominant difficulty drivers correlated with lower success rates are SNR and planet multiplicity measured by Pearson Correlation (Appendix B.3).

Frontier agents fail consistently on real-data subset. The **Real** column of Table 1 reports performance on the 20 tasks based on real-world RV data (Appendix A.3). Across three runs, no model achieves a single pass (0% for all 8 models). Among episodes that produce a submission, ΔBIC passes universally (100%) and RMS passes in 40%, but Match Score remains at 0% and Count passes in only 27%: the closest cases recover correct orbital periods but overestimate semi-amplitudes (Table 6), failing to fully characterise the underlying planets.

All real-data tasks have been solved by human astronomers as the ground-truth are taken from peer-reviewed papers with confirmed, published solutions. As an independent validation step, we re-fit each system's RV data using `RadVel` (Fulton et al., 2018) and verified that the recovered parameters agree with the published values within their reported uncertainties (Appendix A.3). These tasks are therefore *provably solved* by human experts under the same condition, reflecting a critical capability gap frontier agents have yet to reach human researcher level. This also provides evidence against data contamination: although these published papers could have appeared in training corpora, no frontier models were able to achieve a single success on this subset.

4.1. Statistical Fitting vs. Physics Reasoning

Figure 4 decomposes pass rates into the four evaluation criteria. Statistical criteria (ΔBIC and RMS) remain above 70% across all models and tiers, confirming that agents reliably produce well-fitting models. Physical criteria (Match Score and Planet Count), by contrast, drop below 40% on Hard tasks for every model. This gap implies that models optimize within a fixed hypothesis (curve fitting) rather than searching over physically plausible configurations (model selection), which is the core reasoning bottleneck that STARGAZER is designed to expose.

4.2. Test-Time Scaling, Resource Budget and Completion Rate

Test-Time Scaling in Pass@3 We found pass@3 based on 3 independent runs lead to substantial stochasticity: four models reach 95% Pass@3 on Easy-tier despite much lower mean pass rates of 40–80%, and GPT-5.2 attains the highest Hard-tier Pass@3 (12.5%). It's worth noting that the 3 runs are fully independent and do not leak any information/experience from 1 run to another, the improvement here is simply a result of more attempts rather than longer reasoning.

Test-Time Scaling in Single Run. On the other hand, we have a reasoning token budget for each independent run and report how often agents finish within the budget constraints (token limit, timeout, or tool-call cap) in the Env Done columns in Table 1. In hard tasks, six of eight models complete fewer than 6% of tasks naturally, where the majority is cut off by budget limits. GPT-5.3-codex is a notable outlier, completing 7.5% of Hard tasks naturally, driven by its compact output style ($\sim$4K tokens per task).

4.3. Self-Generated Skills.

Following Li et al. (2026), we extract a `skills.md` summary from successful Easy-tier trajectories generated by Opus 4.6 (which is *not* among the eight models evaluated in the main experiment, so the skills document is fully in-

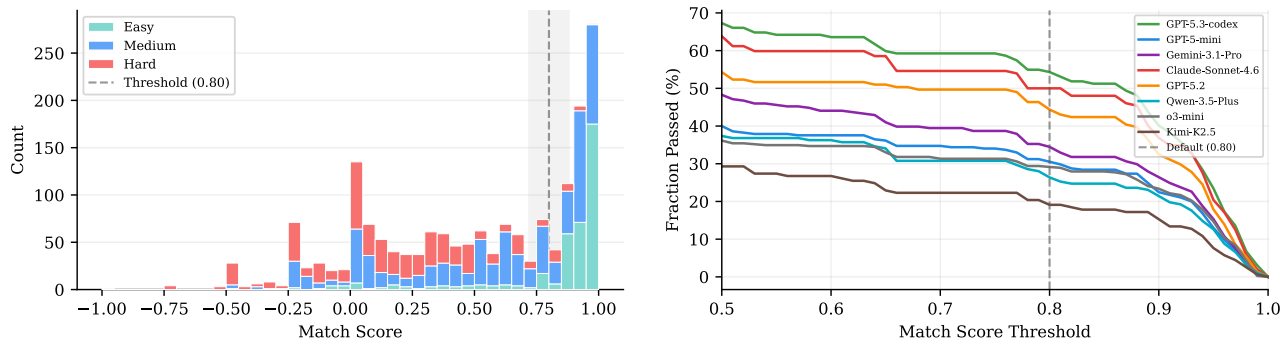

Figure 3. **(a)** Match-score distribution across all submitted episodes colored by difficulty tier. The shaded band marks the $\pm 10\%$ sensitivity region around the default threshold (0.80). **(b)** Pass rate as a function of the match threshold for each model. Rankings are preserved across the entire 0.5–1.0 range. Pass rates are computed as the unweighted fraction across all 100 synthetic tasks; per-tier results in Table 1 remain our primary metric.

	Pass Rate (%)			Env Done (%)			Pass@3 (%)			Real
Model	Easy	Med	Hard	Easy	Med	Hard	Easy	Med	Hard	(20 tasks)
Classical Pipeline	95.0	35.0	5.0	100	100	100	95.0	35.0	5.0	—
Nested Sampling	95.0	32.5	0.0	100	100	100	95.0	32.5	0.0	—
o3-mini	40.0	24.2	0.0	76.7	35.8	4.2	**95.0**	**40.0**	0.0	0.0
GPT-5-mini	76.7	33.6	2.5	76.7	46.7	5.0	**95.0**	35.0	2.5	0.0
GPT-5.2	40.0	30.0	**5.8**	40.0	33.3	5.8	75.0	37.5	**12.5**	0.0
Kimi-K2.5	13.3	17.9	0.8	13.3	19.2	2.5	40.0	35.0	2.5	0.0
Qwen-3.5-Plus	26.7	25.0	1.6	25.0	25.8	1.7	60.0	30.0	2.5	0.0
Gemini-3.1-Pro	71.7	**35.0**	5.0	71.7	35.0	5.8	**95.0**	35.0	7.5	0.0
Claude-Sonnet-4.6	68.3	22.5	0.8	68.3	28.3	0.8	75.0	32.5	2.5	0.0
GPT-5.3-codex	**80.0**	30.8	4.2	**88.3**	**48.3**	**7.5**	**95.0**	**40.0**	7.5	0.0

Table 1. Main results on STARGAZER across difficulty tiers computed over three independent runs. Pass rates and Env Done rates are averaged; Pass@3 reports the fraction of tasks solved in at least one run. **Bold** = best per column; underline = second best.

dependent of the evaluation results). The skill is then provided to 4 tested models at inference time (Table 2), which shows boost to Easy-tier pass tasks for three of four models (up to +28.3 pp). However, an episode-level analysis (Appendix B.5, B.4) reveals that these gains are largely driven by *efficiency* rather than *reasoning*: skills compress the workflow so that previously budget-exceeding episodes now reach the submission stage, although Hard-tier Match Score remains below 33% for all models. For weaker models, skills even degrade Hard-tier RMS, suggesting that rigid templates may have interfered with the more advanced strategies needed for complex systems. The statistical-physical dissociation from Section 4.2 persists across all models.

4.4. Case Studies

We trace two representative trajectories—one success and one failure—that illustrate the behavioral divide between agents that escalate model complexity and those that do not (full step-by-step traces in Appendix E).

✓ Success Case: 2-Planet Recovery *GPT-5.2 | 11 steps*

Task Seed 96 (Medium)

Planets	2 (P = 31.2, 117.6 d)	**Observations**	75
Noise model	White (σ_w = 1.05 m s^{-1})	**SNR**	8.9

Difficulty. Planet 2 dominates the RV signal ($K_2 \approx 30\,\mathrm{m\,s^{-1}}$, $e_2 = 0.25$), but its period (117.6 d) exceeds the data span (106.7 d), causing the periodogram to converge on a biased estimate (~ 105 d). Planet 1 ($P = 31.2$ d, $K_1 \approx 14\,\mathrm{m\,s^{-1}}$) is buried in the residuals.

Steps 1–5 · Protocol & Baseline Reads protocol, computes weighted Lomb–Scargle periodogram (40 000 frequencies). Dominant peak at $P \approx 105$ d (power = 0.80). Baseline one-sine: $P = 104.8$ d, RMS = 10.68 m s^{-1} (RMS/$\tilde{\sigma}$ = 10.1). Gate: **Kepler = YES.**

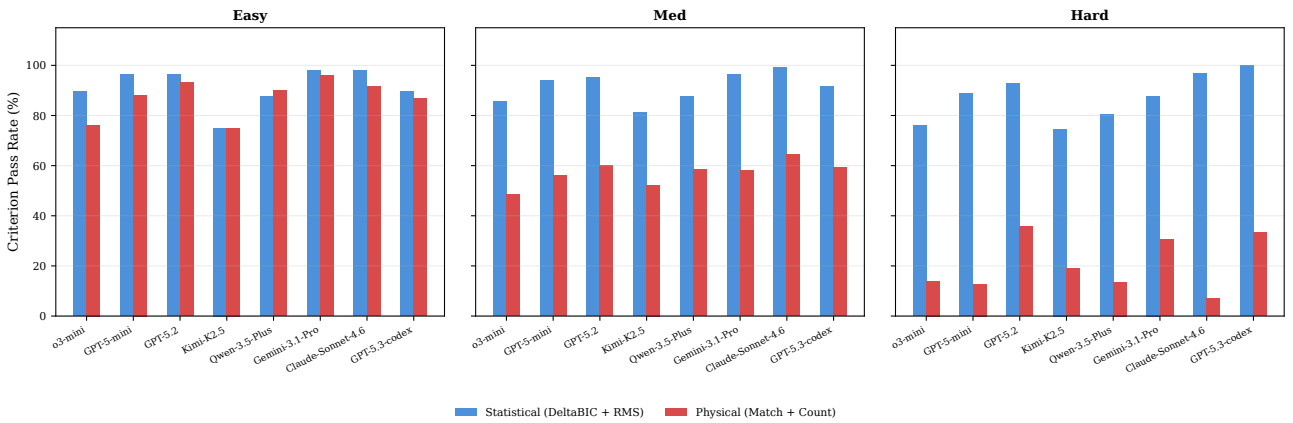

Figure 4. Statistical (blue, mean of ΔBIC and RMS) versus physical (red, mean of Match Score and Planet Count) criterion pass rates by difficulty tier. Statistical pass rates stay high while physical recovery drops from Easy to Hard.

	Pass Rate (%)						Pass@3 (%)					
	Default			+ Skills			Default			+ Skills		
Model	Easy	Med	Hard	Easy	Med	Hard	Easy	Med	Hard	Easy	Med	Hard
GPT-5-mini	76.7	33.6	2.5	$90.0_{+13.3}$	$35.8_{+2.2}$	$2.5_{+0.0}$	95.0	35.0	2.5	$95.0_{+0.0}$	$37.5_{+2.5}$	$2.5_{+0.0}$
Qwen-3.5-Plus	26.7	25.0	1.6	$48.3_{+21.6}$	$23.3_{-1.7}$	$0.0_{-1.6}$	60.0	30.0	2.5	$90.0_{+30.0}$	$35.0_{+5.0}$	$0.0_{-2.5}$
Gemini-3.1-Pro	71.7	35.0	5.0	$100.0_{+28.3}$	$56.7_{+21.7}$	$16.7_{+11.7}$	95.0	35.0	7.5	$100.0_{+5.0}$	$80.0_{+45.0}$	$35.0_{+27.5}$
GPT-5.3-codex	80.0	30.8	4.2	$74.6_{-5.4}$	$53.2_{+22.4}$	$25.6_{+21.4}$	95.0	40.0	7.5	$100.0_{+5.0}$	$67.5_{+27.5}$	$31.6_{+24.1}$

Table 2. Effect of domain-expert skills injection on pass rates (%) and Pass@3 (%), averaged over three independent runs. Colored superscripts show the change in percentage points (+ = improvement, − = degradation).

Steps 6–9 · Iterative Keplerian Fitting 1-planet Keplerian converges to $P = 300$ d at the search boundary (RMS $= 9.16\,\mathrm{m\,s^{-1}}$)—a diagnostic sign that the true period exceeds the data span. Residual periodogram reveals a strong second signal at $P \approx 31.0$ d (power $= 0.84$). 2-planet fit: $P_1 = 116.6$ d, $P_2 = 31.1$ d; RMS drops to $\mathbf{1.13}\,\mathrm{m\,s^{-1}}$ (RMS$/\tilde{\sigma} = 1.07$).

Step 11 · Submit #1 (2 planets) → Pass Match $= 0.932$, RMS $= 1.13\,\mathrm{m\,s^{-1}}$, ΔBIC/pt $= 513$.
✓ bic ✓ rms ✓ match ✓ count

Resources: 61 K input + 7 K output tokens, 0 code errors, 1 submission.

Takeaway. The agent follows a textbook *peel-and-search* strategy: fit the strongest signal, subtract, search residuals for the next. A key diagnostic moment occurs when the 1-planet fit converges to the period upper bound ($P = 300$ d): instead of submitting, the agent recognises incomplete convergence and checks residuals, where the 31 d signal guides the 2-planet fit that simultaneously corrects P_1 to 116.6 d. Both planets are recovered in 11 steps with a single submission, consuming only 68 K tokens.

Successful agents treat failed submissions as informative feedback: they use mismatched diagnostics (e.g., low Match despite good RMS) to revise the hypothesis, escalate model complexity (adding planets), and validate the new fit via residual checks and follow-up periodograms before resubmitting. In contrast, failed agents often lock onto a single hypothesis and fall into repetitive resubmissions that do not incorporate new evidence, exhausting the step and token budget without meaningful exploration.

✗ Failure Case: Stuck in a Local Minimum *GPT-5-mini | 40 steps*

Task Seed `196` (Hard)

Planets	3 (P = 76, 112, 167 d)	**Observations**	43
Noise model	White (σ_w = $0.93\,\mathrm{m\,s^{-1}}$)	**Resonance**	Double (≈ 3:2:1)

Difficulty. The three planets sit in a double-resonance chain with period ratios near 3:2:1, producing alias peaks and combination frequencies in the periodogram that make it easy for an iterative search to lock onto spurious solutions instead of the true orbital periods.

Steps 1–12 · Protocol, Periodogram & Fitting Runs the standard pipeline: protocol, periodogram, baseline, and iterative fitting. Converges on a 2-planet model with alias periods (111.7, 164.5) d that achieve low RMS ($1.08\,\mathrm{m\,s^{-1}}$) but do not correspond to real planets.

Step 13 · Submit #1 (2 planets) → Fail RMS $= 7.71\,\mathrm{m\,s^{-1}}$; recovered parameters are wildly incorrect due to `l_rad` format error.

Steps 14–40 · 9 More Submissions → All Fail By step 20 the format error is fixed and RMS $= 1.08\,\mathrm{m\,s^{-1}}$, but match $= -0.008$: the detected periods are aliases of the true resonant system. **Submissions 3–10 are near-identical**: the same wrong answer is resubmitted without any model escalation.

Resources: 696 K input + 34 K output tokens, 14 code errors, 10 submissions, all failed.

Diagnostics: ✓ bic ✓ rms ✗ match ✗ count (2 vs. 3 true)

Takeaway. The agent converges to a 2-planet model that achieves good RMS but corresponds to *alias periods*, a classic RV pitfall where near-resonant planets produce periodogram peaks at combination frequencies. After reaching this local minimum, the agent cannot escape: it repeats the same submission many times without attempting a 3-planet model. The failure consumes 10.7× more tokens (730 K vs. 68 K) than the success case, demonstrating that *more computation does not compensate for the inability to revise the model hypothesis*.

4.5. Takeaway for Future Model Development

The evaluation of frontier agents within STARGAZER offers several insights for next-generation AI scientists. More broadly, STARGAZER also provides a scalable environment for agentic RL in iterative, feedback-driven scientific workflows:

Goodness-of-fit ≠ physical recovery. Future model developers should pay more attention to the construct validity of their evaluation metric. High scores on statistical goodness-of-fit do not necessarily translate to recovering meaningful physical parameters. Training agents for science should move beyond optimizing for residuals and incorporate metrics that explicitly test scientific validity in the context of the task. More broadly, agent policies should treat diagnostic mismatches as triggers for hypothesis revision and model-complexity escalation, rather than as signals to spend more compute on the same solution.

Procedural scaffolding has limits. On relatively simple tasks, self-generated skills improve efficiency of agentic performance, but they do not fundamentally improve physical reasoning on Hard tasks. This finding is consistent with recent studies on the limitation of skills (Li et al., 2026), where domain expertise is required to curate genuinely generalizable skills. In practice, it may be beneficial to pair procedural scaffolds with robustness mechanisms such as strict output-format compliance and automated harness.

Mind the sim-to-real gap. For agentic RL training, the sim-to-real gap should be treated as a primary consideration. It can determine whether policies learned under dense and well-structured simulated feedback transfer to real RV data that exhibit noise, sparse sampling, and instrument systematics. We therefore suggest evaluating transfer explicitly and designing physics-grounded data mixtures and curricula that disentangle what is controllable in simulation from what must be handled by the agent at deployment time.

5. Conclusion

We introduced STARGAZER, a scalable environment for AI agents on the iterative, multi-step workflow of exoplanet discovery via radial-velocity analysis. By moving beyond static QA toward dynamic, feedback-driven scientific reasoning, STARGAZER exposes capability gaps that existing benchmarks cannot detect. Our evaluation of eight frontier models reveals a consistent pattern: agents are proficient at numerical optimization but struggle with the physical reasoning that distinguishes curve fitting from scientific discovery. This statistical–physical dissociation persists across models, difficulty tiers, and even when agents are equipped with domain-expert skills or additional compute budget. By quantifying these limitations in a physically grounded, infinitely scalable setting, STARGAZER provides a concrete target for developing more capable scientific agents that can not only fit data, but also interpret what the fit means.

Acknowledgements

This work is supported in part by the National Science and Engineering Research Council of Canada, the Dunlap Institute of Astronomy & Astrophysics (seed funding), Anthropic (compute credits), OpenAI (superalignment grant), the German Federal Ministry of Education and Research (BMBF) and Tübingen AI Center (FKZ: 01IS18039B), and the Machine Learning Cluster of Excellence (EXC 2064/1, Project 390727645).

LLM Usage Statement

We used large language models to assist with drafting and revising prose and with minor LaTeX editing, and an image-generation model to create the robot icons used in Figures 1 and 2. All technical content, experimental design, code, analyses, and results were produced and verified by the authors.

Impact Statement

STARGAZER is introduced to help researchers evaluate AI agents on iterative, physics-grounded model-fitting workflows. By providing a standardized and reproducible benchmark for radial-velocity inference, it enables clearer comparisons across models, tool-use strategies, feedback policies, and domain-scientific scaffolds. We hope this benchmark will support the development of scientific agents that can not only optimize numerical fits, but also recover physically meaningful explanations from noisy data.

We also encourage careful use of this benchmark in deployment-focused settings. STARGAZER is not intended to certify autonomous discovery systems for unsupervised scientific deployment. Over-optimizing agents for benchmark feedback may encourage brittle strategies that exploit simulator assumptions, statistical thresholds, or repeated submissions rather than robust scientific reasoning. In real astronomical workflows, incorrect model fits can mislead follow-up observations and downstream interpretation. The research community therefore shares a responsibility to report results transparently, verify evidence against independent physical criteria, and evaluate transfer from simulation to real archival data before drawing deployment-oriented conclusions. This work uses simulated and publicly available archival radial-velocity data and does not involve human subjects or personal data.

References

Anglada-Escudé, G., López-Morales, M., and Chambers, J. E. How eccentric orbital solutions can hide planetary systems in 2:1 resonant orbits. *The Astrophysical Journal*, 709(1):168–178, December 2009. ISSN 1538-4357. doi: 10.1088/0004-637x/709/1/168. URL http://dx.doi.org/10.1088/0004-637X/709/1/168.

Benedict, G. F., McArthur, B. E., Nelan, E. P., Wittenmyer, R., Barnes, R., Smotherman, H., and Horner, J. The μ Arae planetary system: Radial velocities and astrometry. *The Astronomical Journal*, 163:295, 2022. doi: 10.3847/1538-3881/ac6ac8.

Birkby, J. L., de Kok, R. J., Brogi, M., Schwarz, H., and Snellen, I. A. G. Discovery of water at high spectral resolution in the atmosphere of 51 Peg b. *The Astronomical Journal*, 153:138, 2017. doi: 10.3847/1538-3881/aa5c87.

Boisse, I., Moutou, C., Vidal-Madjar, A., Bouchy, F., Pont, F., Hébrard, G., Bonfils, X., Croll, B., Delfosse, X., Desort, M., Forveille, T., Lagrange, A.-M., Loeillet, B., Lovis, C., Matthews, J. M., Mayor, M., Pepe, F., Perrier, C., Queloz, D., Rowe, J. F., Santos, N. C., Ségransan, D., and Udry, S. Stellar activity of planetary host star hd 189733. *Astronomy & Astrophysics*, 495(3):959–966, January 2009. ISSN 1432-0746. doi: 10.1051/0004-6361:200810648. URL http://dx.doi.org/10.1051/0004-6361:200810648.

Brenner, M. P., Cohen-Addad, V., and Woodruff, D. P. Solving an open problem in theoretical physics using AI-assisted discovery. *CoRR*, abs/2603.04735, 2026. URL https://arxiv.org/abs/2603.04735.

Budavári, T. and Basu, A. Probabilistic cross-identification in crowded fields as an assignment problem. *The Astronomical Journal*, 152(4):86, September 2016. ISSN 1538-3881. doi: 10.3847/0004-6256/152/4/86. URL http://dx.doi.org/10.3847/0004-6256/152/4/86.

Butler, R. P., Wright, J. T., Marcy, G. W., Fischer, D. A., Vogt, S. S., Tinney, C. G., Jones, H. R. A., Carter, B. D., Johnson, J. A., McCarthy, C., and Penny, A. J. Catalog of nearby exoplanets. *The Astrophysical Journal*, 646: 505–522, 2006. doi: 10.1086/504701.

Cao, Y., Lai, S., Huang, J., Zhang, Y., Lawrence, Z., Bhakta, R., Thomas, I. F., Cao, M., Tsai, C.-H., Zhou, Z., Zhao, Y., Liu, H., Marinoni, A., Arefiev, A., and Yu, R. Simulcost: A cost-aware benchmark and toolkit for automating physics simulations with llms, 2026. URL https://arxiv.org/abs/2603.20253.

Center for AI Safety, Scale AI, and HLE Contributors Consortium. A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, 649:1139–1146, 2026. doi: 10.1038/s41586-025-09962-4. URL https://www.nature.com/articles/s41586-025-09962-4.

Chen, Z., Chen, S., Ning, Y., Zhang, Q., Wang, B., Yu, B., Li, Y., Liao, Z., Wei, C., Lu, Z., Dey, V., Xue, M., Baker, F. N., Burns, B., Adu-Ampratwum, D., Huang, X., Ning, X., Gao, S., Su, Y., and Sun, H. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery, 2024. URL https://arxiv.org/abs/2410.05080.

Chen, Z., Chen, S., Ning, Y., Zhang, Q., Wang, B., Yu, B., Li, Y., Liao, Z., Wei, C., Lu, Z., Dey, V., Xue, M., Baker, F. N., Burns, B., Adu-Ampratwum, D., Huang, X., Ning, X., Gao, S., Su, Y., and Sun, H. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery, 2025. URL https://arxiv.org/abs/2410.05080.

Correia, A. C. M., Udry, S., Mayor, M., Benz, W., Bertaux, J., Bouchy, F., Laskar, J., Lovis, C., Mordasini, C., Pepe, F. A., and Queloz, D. The harps search for southern extra-solar planets - xvi. hd 45364, a pair of planets in a 3:2 mean motion resonance. *Astronomy and Astrophysics*, 496:521–526, 2009. URL https://api.semanticscholar.org/CorpusID:119235349.

Cumming, A. Detectability of extrasolar planets in radial velocity surveys. *Monthly Notices of the Royal Astronomical Society*, 354:1165–1176, 2004.

Delisle, J.-B., Ségransan, D., Dumusque, X., et al. The HARPS search for southern extra-solar planets. XLIII. A compact system of four super-earth planets orbiting HD 215152. *Astronomy & Astrophysics*, 614:A133, 2018. doi: 10.1051/0004-6361/201732529.

Fulton, B. J., Petigura, E. A., Blunt, S., and Sinukoff, E. RadVel: The radial velocity modeling toolkit. *Publications of the Astronomical Society of the Pacific*, 130(986): 044504, 2018. doi: 10.1088/1538-3873/aaaa70.

Haywood, R. D., Collier Cameron, A., Queloz, D., Barros, S. C. C., Deleuil, M., Fares, R., Gillon, M., Lanza, A. F., Lovis, C., Moutou, C., Pepe, F., Pollacco, D., Santerne, A., Ségransan, D., and Unruh, Y. C. Planets and stellar activity: hide and seek in the corot-7 system. *Monthly Notices of the Royal Astronomical Society*, 443(3):2517–2531, July 2014. ISSN 0035-8711. doi: 10.1093/mnras/stu1320. URL http://dx.doi.org/10.1093/mnras/stu1320.

He, C., Luo, R., Bai, Y., Hu, S., Thai, Z. L., Shen, J., Hu, J., Han, X., Huang, Y., Zhang, Y., Liu, J., Qi, L., Liu, Z., and Sun, M. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pp. 3828–3850, 2024. URL https://arxiv.org/abs/2402.14008.

Hopkins, A. M., Whiting, M. T., Seymour, N., Chow, K. E., Norris, R. P., Bonavera, L., Breton, R., Carbone, D., Ferrari, C., Franzen, T. M. O., Garsden, H., González-Nuevo, J., Hales, C. A., Hancock, P. J., Heald, G., Herranz, D., Huynh, M., Jurek, R. J., López-Caniego, M., Massardi, M., Mohan, N., Molinari, S., Orrù, E., Paladino, R., Pestalozzi, M., Pizzo, R., Rafferty, D., Röttgering, H. J. A., Rudnick, L., Schisano, E., Shulevski, A., Swinbank, J., Taylor, R., and van der Horst, A. J. The askap/emu source finding data challenge. *Publications of the Astronomical Society of Australia*, 32, 2015. ISSN 1448-6083. doi: 10.1017/pasa.2015.37. URL http://dx.doi.org/10.1017/pasa.2015.37.

Joseph, S. A., Husain, S. M., Offner, S. S. R., Juneau, S., Torrey, P., Bolton, A. S., Farias, J. P., Gaffney, N., Durrett, G., and Li, J. J. AstroVisBench: A code benchmark for scientific computing and visualization in astronomy. In *Advances in Neural Information Processing Systems 38, NeurIPS 2025, Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=qXiTFAgEx4.

Kipping, D. M. Parametrizing the exoplanet eccentricity distribution with the Beta distribution. *Monthly Notices of the Royal Astronomical Society: Letters*, 434:L51–L55, 2013.

Koblischke, N., Jang, H., Menou, K., and Ali-Dib, M. Gravity-bench-v1: A benchmark on gravitational physics discovery for agents. *arXiv preprint arXiv:2501.18411*, 2025. URL https://arxiv.org/abs/2501.18411.

Krenn, M., Pollice, R., Guo, S. Y., Aldeghi, M., Cervera-Lierta, A., Friederich, P., dos Passos Gomes, G., Häse, F., Jinich, A., Nigam, A., Yao, Z., and Aspuru-Guzik, A. On scientific understanding with artificial intelligence. *Nature Reviews Physics*, 4(12):761–769, 2022. URL https://doi.org/10.1038/s42254-022-00518-3.

Kuhn, H. W. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2): 83–97, 1955. doi: 10.1002/nav.3800020109.

Laughlin, G., Marcy, G. W., Vogt, S. S., Fischer, D. A., and Butler, R. P. On the eccentricity of HD 209458b. *The Astrophysical Journal Letters*, 629:L121–L124, 2005. doi: 10.1086/444558.

Li, X., Chen, W., Liu, Y., Zheng, S., Chen, X., He, Y., Li, Y., You, B., Shen, H., Sun, J., Wang, S., Li, B., Zeng, Q., Wang, D., Zhao, X., Wang, Y., Chaim, R. B., Di, Z., Gao, Y., He, J., He, Y., Jing, L., Kong, L., Lan, X., Li, J., Li, S., Li, Y., Lin, Y., Liu, X., Liu, X., Lyu, H., Ma, Z., Wang, B., Wang, R., Wang, T., Ye, W., Zhang, Y., Xing, H., Xue, Y., Dillmann, S., and chung Lee, H. Skillsbench: Benchmarking how well agent skills work across diverse tasks, 2026. URL https://arxiv.org/abs/2602.12670.

Lovis, C., Mayor, M., Pepe, F., Alibert, Y., Benz, W., Bouchy, F., Correia, A. C. M., Laskar, J., Mordasini, C., Queloz, D., Santos, N. C., Udry, S., Bertaux, J.-L., and Sivan, J.-P. An extrasolar planetary system with three Neptune-mass planets. *Nature*, 441:305–309, 2006a. doi: 10.1038/nature04828.

Lovis, C., Mayor, M., Pepe, F., Alibert, Y., Benz, W., Bouchy, F., Correia, A. C. M., Laskar, J., Mordasini, C., Queloz, D., Santos, N. C., Udry, S., Bertaux, J.-L., and Sivan, J.-P. An extrasolar planetary system with three neptune-mass planets. *Nature*, 441(7091):305–309, May 2006b. ISSN 1476-4687. doi: 10.1038/nature04828. URL http://dx.doi.org/10.1038/nature04828.

Lovis, C., Ségransan, D., Mayor, M., Udry, S., Benz, W., Bertaux, J.-L., Bouchy, F., Correia, A. C. M., Laskar, J., Lo Curto, G., Mordasini, C., Pepe, F., Queloz, D., and Santos, N. C. The harps search for southern extra-solar planets: Xxviii. up to seven planets orbiting hd 10180: probing the architecture of low-mass planetary systems. *Astronomy & Astrophysics*, 528:A112, March 2011. ISSN 1432-0746. doi: 10.1051/0004-6361/201015577. URL http://dx.doi.org/10.1051/0004-6361/201015577.

Lupidi, A., Gauri, B., Foster, T. S., Omari, B. A., Magka, D., Pepe, A., Audran-Reiss, A., Aghamelu, M., Baldwin, N., Cipolina-Kun, L., Gagnon-Audet, J.-C., Leow, C. H., Lefdal, S., Mossalam, H., Moudgil, A., Nazir, S., Tewolde, E., Urrego, I., Estape, J. A., Budhiraja, A., Chaurasia, G., Charnalia, A., Dunfield, D., Hambardzumyan, K., Izcovich, D., Josifoski, M., Mediratta, I., Niu, K., Pathak, P., Shvartsman, M., Toledo, E., Protopopov, A., Raileanu, R., Miller, A., Shavrina, T., Foerster, J., and Bachrach, Y. Airs-bench: a suite of tasks for frontier ai research science agents, 2026. URL https://arxiv.org/abs/2602.06855.

Majumder, B. P., Surana, H., Agarwal, D., Mishra, B. D., Meena, A., Prakhar, A., Vora, T., Khot, T., Sabharwal,

A., and Clark, P. DiscoveryBench: Towards data-driven discovery with large language models, 2024. URL https://arxiv.org/abs/2407.01725.

Mak, C.-W., Zhu, G., Zhang, B., Li, H., Chi, X., Zhang, K., Wu, Y., He, Y., Fan, C.-K., Lu, W., Ge, K., Fang, X., He, H., Lu, K., Xu, T., Zhang, L., Ni, Y., Li, Y., and Zhang, S. Physicsmind: Sim and real mechanics benchmarking for physical reasoning and prediction in foundational vlms and world models, 2026. URL https://arxiv.org/abs/2601.16007.

Mayor, M. and Queloz, D. A jupiter-mass companion to a solar-type star. *Nature*, 378:355–359, 1995. URL https://doi.org/10.1038/378355a0.

Mayor, M., Bonfils, X., Forveille, T., Delfosse, X., Udry, S., Bertaux, J.-L., Beust, H., Bouchy, F., Lovis, C., Pepe, F., Perrier, C., Queloz, D., and Santos, N. C. The HARPS search for southern extra-solar planets. XVIII. an Earth-mass planet in the GJ 581 planetary system. *Astronomy & Astrophysics*, 507:487–494, 2009a. doi: 10.1051/0004-6361/200912172.

Mayor, M., Udry, S., Lovis, C., Pepe, F., Queloz, D., Benz, W., Bertaux, J.-L., Bouchy, F., Mordasini, C., and Segransan, D. The HARPS search for southern extra-solar planets. XIII. A planetary system with 3 super-Earths. *Astronomy & Astrophysics*, 493:639–644, 2009b. doi: 10.1051/0004-6361:200810451.

Naef, D., Mayor, M., Beuzit, J. L., Perrier, C., Queloz, D., Sivan, J. P., and Udry, S. The ELODIE survey for northern extra-solar planets. III. Three planetary candidates detected with ELODIE. *Astronomy & Astrophysics*, 414: 351–359, 2004. doi: 10.1051/0004-6361:20034091.

Pepe, F., Lovis, C., Ségransan, D., Benz, W., Bouchy, F., Bouchy, F., Dumusque, X., Mayor, M., Queloz, D., Santos, N. C., and Udry, S. The harps search for earth-like planets in the habitable zone - i. very low-mass planets around hd 20794, hd 85512, and hd 192310. *Astronomy and Astrophysics*, 534:16, 2011. URL https://api.semanticscholar.org/CorpusID:15088852.

Perryman, M. *The Exoplanet Handbook*. Cambridge University Press, 2nd edition, 2018. doi: 10.1017/9781108304160. URL https://www.cambridge.org/core/books/exoplanet-handbook/750759E015FDCF469D141F0046198519.

Queloz, D., Henry, G. W., Sivan, J. P., Baliunas, S. L., Beuzit, J.-L., Donahue, R. A., Mayor, M., Naef, D., Perrier, C., and Udry, S. No planet for HD 166435. *Astronomy & Astrophysics*, 379:L5–L8, 2001.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. GPQA: A graduate-level google-proof Q&A benchmark. In *First Conference on Language Modeling, COLM 2024*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Rein, H. and Liu, S.-F. REBOUND: an open-source multi-purpose N-body code for collisional dynamics. *Astronomy & Astrophysics*, 537:A128, 2012. doi: 10.1051/0004-6361/201118085.

Rivera, E. J., Laughlin, G., Butler, R. P., Vogt, S. S., Haghighipour, N., and Meschiari, S. The Lick-Carnegie exoplanet survey: a Uranus-mass fourth planet for GJ 876 in an extrasolar Laplace configuration. *The Astrophysical Journal*, 719:890–899, 2010. doi: 10.1088/0004-637X/719/1/890.

Shen, Y., Yang, Y., Xi, Z., Hu, B., Sha, H., Zhang, J., Peng, Q., Shang, J., Huang, J., Fan, Y., Tong, J., Dou, S., Zhang, M., Bai, L., Yin, Z., Gui, T., Ma, X., Zhang, Q., Huang, X., and Jiang, Y.-G. Sciagentgym: Benchmarking multi-step scientific tool-use in llm agents, 2026. URL https://arxiv.org/abs/2602.12984.

Shi, J., Tang, X., Huang, Y., Li, Y., Kong, X., Zhang, Y., and Yue, C. Astrommbench: A benchmark for evaluating multimodal large language models capabilities in astronomy, 2025. URL https://arxiv.org/abs/2510.00063.

Siegel, Z. S., Kapoor, S., Nagdir, N., Stroebl, B., and Narayanan, A. CORE-Bench: Fostering the credibility of published research through a computational reproducibility agent benchmark, 2024. URL https://arxiv.org/abs/2409.11363.

Tian, M., Gao, L., Zhang, S. D., Chen, X., Fan, C., Guo, X., Haas, R., Ji, P., Krongchon, K., Li, Y., Liu, S., Luo, D., Ma, Y., Tong, H., Trinh, K., Tian, C., Wang, Z., Wu, B., Yin, S., Zhu, M., Lieret, K., Lu, Y., Liu, G., Du, Y., Tao, T., Press, O., Callan, J., Huerta, E. A., and Peng, H. Scicode: A research coding benchmark curated by scientists. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/36850592258c8c41cecdaa3dea5ff7de-Abstract-Datasets_and_Benchmarks_Track.html.

Ting, Y.-S., Nguyen, T. D., Ghosal, T., Pan, R., Arora, H., Sun, Z., de Haan, T., Ramachandra, N., Wells, A.,

Madireddy, S., and Accomazzi, A. Astromlab 1: Who wins astronomy jeopardy!?, 2024. URL https://arxiv.org/abs/2407.11194.

Vogt, S. S., Butler, R. P., Marcy, G. W., Fischer, D. A., Henry, G. W., Laughlin, G., Wright, J. T., and Johnson, J. A. Five new multicomponent planetary systems. *The Astrophysical Journal*, 632:638–658, 2005. doi: 10.1086/432901.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., Anandkumar, A., Bergen, K., Gomes, C. P., Ho, S., Kohli, P., Lasenby, J., Leskovec, J., Liu, T., Manrai, A., Marks, D., Ramsundar, B., Song, L., Sun, J., Tang, J., Veličković, P., Welling, M., Zhang, L., Coley, C. W., Bengio, Y., and Zitnik, M. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023. URL https://doi.org/10.1038/s41586-023-06221-2.

Wang, W., Chen, X., Gong, J., Huang, X., and Qiu, X. Astroreason-bench: Evaluating unified agentic planning across heterogeneous space planning problems, 2026. URL https://arxiv.org/abs/2601.11354.

Wang, X., Hu, Z., Lu, P., Zhu, Y., Zhang, J., Subramaniam, S., Loomba, A. R., Zhang, S., Sun, Y., and Wang, W. SciBench: Evaluating college-level scientific problem-solving abilities of large language models. In *Proceedings of the 41st International Conference on Machine Learning, ICML 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 50622–50649. PMLR, 2024. URL https://proceedings.mlr.press/v235/wang24z.html.

Winn, J. N. and Fabrycky, D. C. The occurrence and architecture of exoplanetary systems. *Annual Review of Astronomy and Astrophysics*, 53:409–447, 2015. doi: 10.1146/annurev-astro-082214-122246. URL https://ui.adsabs.harvard.edu/abs/2015ARA&A..53..409W.

Wright, J. T., Upadhyay, S., Marcy, G. W., Fischer, D. A., Ford, E. B., and Johnson, J. A. Ten new and updated multiplanet systems and a survey of exoplanetary systems. *The Astrophysical Journal*, 693:1084–1099, 2009. doi: 10.1088/0004-637X/693/2/1084.

Xiang, K., Li, H., Zhang, T. J., Huang, Y., Liu, Z., Qu, P., He, J., Chen, J., Yuan, Y., Han, J., Xu, H., Li, H., Sachan, M., and Liang, X. Seephys: Does seeing help thinking? – benchmarking vision-based physics reasoning. In *Advances in Neural Information Processing Systems 38, NeurIPS 2025, Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=APNWmytTCS.

Xu, X., Xu, Q., Xiao, T., Chen, T., Yan, Y., Zhang, J., Diao, S., Yang, C., and Wang, Y. UGPhysics: A comprehensive benchmark for undergraduate physics reasoning with large language models. In *Proceedings of the 42nd International Conference on Machine Learning, ICML 2025*, Proceedings of Machine Learning Research. PMLR, 2025. URL https://icml.cc/virtual/2025/poster/45927.

Zhang, X., Dong, Y., Wu, Y., Huang, J., Jia, C., Fernando, B., Shou, M. Z., Zhang, L., and Liu, J. PhysReason: A comprehensive benchmark towards physics-based reasoning, 2025. URL https://arxiv.org/abs/2502.12054.

Zhao, W., Ma, Q., Shi, J., Wu, S., Han, J., Xiao, Y., Chen, S.-Y., Luo, X., Schmidt, L., and Zou, J. PRISM-physics: Causal DAG-based process evaluation for physics reasoning. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2510.03185.

Zheng, T., Tam, K. K.-W., Nguyen, N. H.-N. K., Xu, B., Wang, Z., Cheng, J., Tsang, H. T., Wang, W., Bai, J., Fang, T., Song, Y., Wong, G. Y., and See, S. Newtonbench: Benchmarking generalizable scientific law discovery in llm agents, 2026. URL https://arxiv.org/abs/2510.07172.

Zhu, M., Tian, M., Yang, X., Zhou, T., Yuan, L., Zhu, P., Chertkov, E., Liu, S., Du, Y., Ji, Z., Das, I., Cao, J., Yu, J., Wu, P., He, J., Su, Y., Jiang, Y., Zhang, Y., Liu, C., Huang, Z., Jia, W., Wang, Y., Jafarpour, F., Zhao, Y., Chen, X., Shelton, J., Young, A. W., Bartolotta, J., Xu, W., Sun, Y., Chu, A., Colussi, V., Akers, C., Brooks, N., Fu, W., Zhao, J., Qi, M., Mu, A., Yang, Y., Zang, A., Lyu, Y., Mai, P., Wilson, C., Guo, X., Zhou, J., Inafuku, D., Xue, C., Gao, L., Yang, Z., Hein, Y., Kahn, Y., Zhou, K., Luo, D., Wilson, J. D., Reilly, J. T., Bandak, D., Press, O., Yang, L., Wang, X., Tong, H., Chia, N., Huerta, E., and Peng, H. Probing the critical point (CritPt) of AI reasoning: a frontier physics research benchmark. *CoRR*, abs/2509.26574, 2025. URL https://arxiv.org/abs/2509.26574.

Appendix

A. Task Construction Details

A.1. Difficulty Scoring Rubric

Each synthetic task is assigned an integer difficulty level $d \in [1, 10]$ based on six physically motivated factors. The difficulty score is computed as

$$d = \mathrm{clip}\big(d_{\mathrm{base}} + d_{\mathrm{SNR}} + d_{\mathrm{res}} + d_{\mathrm{cov}} + d_{\mathrm{obs}} + d_{\mathrm{GP}},\ 1,\ 10\big), \tag{4}$$

where each component is defined in Table 3.

Factor	Condition	Score
Planet multiplicity (d_{base})	1 planet	+1
	2 planets	+2
	3 planets	+3
	4+ planets	+4
Signal-to-noise ratio (d_{SNR})	SNR > 5	0
	SNR > 2	+1
	SNR > 1	+2
	SNR ≤ 1	+3
Resonant configurations (d_{res})	0 resonances	0
	$\geq$1 resonance	$+\min(2,\ n_{\mathrm{res}})$
Coverage of inner period (d_{cov})	$T_{\mathrm{base}}/P_{\mathrm{inner}} \geq 3$	0
	≥ 2	+1
	< 2	+2
Number of observations (d_{obs})	$n_{\mathrm{obs}} \geq 80$	0
	≥ 50	+1
	≥ 30	+2
	< 30	+3
Correlated noise (d_{GP})	No GP	0
	$\sigma_{\mathrm{GP}} < 0.5$	+1
	$\sigma_{\mathrm{GP}} < 1.0$	+2
	$\sigma_{\mathrm{GP}} \geq 1.0$	+3

Table 3. Difficulty scoring rubric. Each factor contributes an additive term to the total difficulty score, which is clipped to $[1, 10]$.

A.2. Synthetic Task Generation Details

Each synthetic task is fully determined by a single random seed. The generation pipeline draws parameters from the following priors:

- **Number of planets:** 1–4, with higher counts at higher difficulty levels.
- **Orbital periods:** log-uniform from $[2, 300]$ days, with a 25% probability of inserting near-resonant pairs (period ratios within 3% of 2:1, 3:2, or 5:3).
- **Minimum masses:** $m \sin i$ drawn from $[0.01, 1.0]\ M_{\mathrm{Jup}}$.
- **Eccentricities:** Kipping Beta distribution (Kipping, 2013) with $\alpha = 0.867$, $\beta = 3.03$.
- **Angular parameters:** argument of periastron ω, longitude of ascending node Ω, and mean longitude ℓ each drawn uniformly from $[0, 2\pi)$.
- **White noise:** measurement uncertainty σ_w drawn from $10^{U(-0.3,\,0.7)}\ \mathrm{m\,s^{-1}}$ (approximately 0.5–5 $\mathrm{m\,s^{-1}}$), with an optional jitter term σ_j added in quadrature.

- **Correlated stellar noise:** included with 40% probability. Modeled as a Gaussian Process with a quasi-periodic rotation kernel (celerite2 `RotationTerm`), parameterized by amplitude $\sigma_{\rm GP} \in [0.05, 1.6]\,{\rm m\,s^{-1}}$ and stellar rotation period $\in [10, 45]$ days.

- **Observation schedule:** timestamps drawn uniformly over a baseline spanning 2–4$\times$ the shortest planetary period, producing an irregularly sampled time grid typical of ground-based surveys. The number of observations ranges from 30 to 100.

The clean RV signal is computed as a multi-Keplerian superposition via numerical solution of Kepler's equation (Newton iteration). Per-instrument systemic velocity offsets γ_i are added for multi-instrument tasks.

Tasks are grouped into three tiers: **Easy** (difficulty 1–2, 20 tasks), **Medium** (difficulty 3–6, 40 tasks), and **Hard** (difficulty 7–10, 40 tasks). Easy tasks are intentionally fewer because low-complexity single-planet systems occupy a smaller region of the physically plausible parameter space. The scoring rubric was calibrated through pilot experiments in two steps. We first assigned provisional weights by reverse-engineering which physical factors most reliably destabilise the standard RV workflow used by human analysts: low SNR, multiplicity, resonances, poor coverage, limited observations, and correlated stellar noise. We then adjusted the weights and tier cutoffs so that pilot pass rates decreased monotonically with nominal difficulty, while keeping the score interpretable as a sum of physically meaningful failure drivers rather than a purely empirical hardness label. Candidate instances that were physically non-identifiable under the realised cadence and noise draw were removed during benchmark construction, so difficulty is calibrated over a pool of tasks intended to be challenging but still solvable in principle.

A.3. Real-World RV Dataset

In addition to the 100 synthetic tasks, STARGAZER includes 20 tasks constructed from published radial velocity datasets of confirmed exoplanetary systems. These tasks span the full range of RV analysis complexity: from single hot Jupiters with $K > 100\,{\rm m\,s^{-1}}$ to multi-planet systems with sub-m s^{-1} signals buried in correlated noise. All identifying information (target names, instrument names, literature references, prior knowledge of planetary parameters) is removed from the task files presented to the agent; the agent receives only time-series data, measurement uncertainties, instrument labels (anonymised as `inst_A`, `inst_B`, ...), and the host star mass.

Table 4 lists the 20 real-data tasks with their provenance. The systems are ordered by estimated analysis difficulty, which reflects the number of planets, signal-to-noise ratio, orbital architecture (resonances, high eccentricity), and data complexity (number of instruments, stellar activity).

Data provenance. Radial velocity time series were obtained from two primary sources: the NASA Exoplanet Archive[1] (Butler et al., 2006; Wright et al., 2009; Lovis et al., 2006b) and VizieR[2] (Naef et al., 2004; Mayor et al., 2009b; Lovis et al., 2011). Table 5 lists the exact archive identifier for each system. Ground-truth orbital parameters are taken from the discovery or characterisation papers listed in the Reference column. For multi-instrument datasets, we preserve the original per-instrument RV zero-point offsets; the agent must independently determine and fit these offsets.

Anonymisation protocol. To prevent data contamination from LLM training corpora, each task is assigned an opaque identifier (`real_001` through `real_020`). All metadata that could reveal the target identity — star name, instrument names, literature references, known orbital parameters, and descriptive text — is stripped from the task file. Instrument labels are replaced with generic identifiers (`inst_A`, `inst_B`, ...). The agent receives only: (i) the time series (t_i, v_i, σ_i), (ii) anonymised instrument labels, and (iii) the host star mass $M_\star$.

Ground truth and evaluation. Real-data tasks are evaluated with the same four criteria as synthetic tasks (§3.3). Ground-truth parameters are taken from the peer-reviewed papers listed in Table 4; we independently verified literature consistency by re-fitting each system with `RadVel`. For contested detections (e.g., GJ 581 d/g), we adopt the most widely accepted published solution.

[1] https://exoplanetarchive.ipac.caltech.edu/
[2] https://vizier.cds.unistra.fr/

ID	System	$N_{\rm obs}$	$N_{\rm pl}$	$K_{\rm max}$ (m s^{-1})	$N_{\rm inst}$	Key Challenge	Reference
Single-planet systems							
real_012	HD 209458	85	1	84.3	1	Rossiter–McLaughlin contamination	Laughlin et al. (2005)
real_001	51 Peg	639	1	56.0	6	Multi-instrument offsets (6 spectrographs)	Birkby et al. (2017)
real_010	HD 189733	33	1	205	1	Active star, sparse data	Boisse et al. (2009)
real_009	HD 179949	88	1	112.6	2	Hot Jupiter, 2 instruments	Butler et al. (2006)
real_014	HD 217107	207	1	139.7	2	Moderate eccentricity, 2 instruments	Wright et al. (2009)
real_020	HD 88133	21	1	35.7	1	Sparse data (21 points)	Butler et al. (2006)
Two-planet systems							
real_007	HD 12661	106	2	74.4	2	Clear period separation	Wright et al. (2009)
real_015	HD 37124	52	2	28.5	1	Long-period outer planet	Vogt et al. (2005)
real_019	HD 74156	95	2	125.0	2	High eccentricities ($e > 0.5$)	Naef et al. (2004)
real_017	HD 45364	58	2	21.2	1	3:2 mean-motion resonance	Correia et al. (2009)
real_013	HD 215152	373	2	0.87	2	Sub-m s^{-1} signals, instrument offset	Delisle et al. (2018)
Three-planet systems							
real_018	HD 69830	74	3	3.51	1	All $K < 4$ m s^{-1}	Lovis et al. (2006a)
real_016	HD 40307	129	3	2.54	1	All $K < 3$ m s^{-1}, close periods	Mayor et al. (2009b)
real_011	HD 20794	187	3	0.85	1	Sub-m s^{-1} signals	Pepe et al. (2011)
Four-or-more-planet systems							
real_004	GJ 876	162	4	214.0	1	Laplace resonance chain	Rivera et al. (2010)
real_008	HD 160691	380	4	37.8	3	3-instrument compilation, wide K range	Benedict et al. (2022)
real_002	55 Cnc	48	2†	71.3	1	Complex architecture (5 planets known)	Naef et al. (2004)
real_005	HD 10180	190	5	4.54	1	5 planets, low K, close periods	Lovis et al. (2011)
real_006	HD 10180 (full)	190	7	4.54	1	7 planets incl. sub-m s^{-1} signals	Lovis et al. (2011)
real_003	GJ 581	119	4	12.5	1	Contested planets, stellar activity	Mayor et al. (2009a)

†The Naef et al. (2004) ELODIE dataset for 55 Cnc contains 48 observations sufficient to constrain 2 of the 5 known planets.

Table 4. Real-world RV tasks included in STARGAZER. $N_{\rm obs}$ is the number of observations; $N_{\rm pl}$ is the number of known planets in the ground-truth model; $K_{\rm max}$ is the semi-amplitude of the dominant planet; $N_{\rm inst}$ is the number of instruments. Tasks are presented to the agent with anonymised identifiers and no prior information about the planetary system.

ID	System	Archive	Identifier / Catalogue
real_001	51 Peg	VizieR	`J/AJ/153/138`
real_002	55 Cnc	VizieR	`J/A+A/414/351` (table3)
real_003	GJ 581	VizieR	`J/A+A/507/487` (table1)
real_004	GJ 876	VizieR	`J/ApJ/719/890` (table1)
real_005	HD 10180 (5p)	VizieR	`J/A+A/528/A112` (table1)
real_006	HD 10180 (7p)	VizieR	`J/A+A/528/A112` (table1)
real_007	HD 12661	NASA Exoplanet Archive	UID 0009683
real_008	HD 160691	VizieR	`J/AJ/163/295` (table2)
real_009	HD 179949	NASA Exoplanet Archive	UID 0094645
real_010	HD 189733	VizieR	`J/A+A/495/959` (table1)
real_011	HD 20794	VizieR	`J/A+A/534/A58` (table1)
real_012	HD 209458	NASA Exoplanet Archive	UID 0108859
real_013	HD 215152	VizieR	`J/A+A/614/A133` (harps_a, harps_b)
real_014	HD 217107	NASA Exoplanet Archive	UID 0113421
real_015	HD 37124	NASA Exoplanet Archive	UID 0026381
real_016	HD 40307	VizieR	`J/A+A/493/639` (table1)
real_017	HD 45364	VizieR	`J/A+A/496/521` (hd45364)
real_018	HD 69830	NASA Exoplanet Archive	UID 0040693
real_019	HD 74156	VizieR	`J/A+A/414/351` (table2)
real_020	HD 88133	NASA Exoplanet Archive	UID 0049813

Table 5. Data provenance for all 20 real-world RV tasks. NASA Exoplanet Archive entries are identified by their unique dataset ID (UID); VizieR entries are identified by their catalogue designation.

Special case: GJ 876 (real_004). GJ 876 hosts a four-planet Laplace resonance chain whose strong planet–planet interactions invalidate the Keplerian superposition assumption; its ground truth is taken from the N-body solution of Rivera et al. (2010). We retain this system intentionally: it tests whether agents can recognise when the standard fitting workflow breaks down. In practice, several agents note that a dominant $\sim$61 d signal leaves a stubborn $\sim$30 d residual, and some speculate about resonance or interactions, but none diagnoses the model family itself as misspecified—they continue escalating within the Keplerian search loop rather than switching to a dynamical model.

Table 6 summarises the five agent submissions that came closest to the published real-data solutions under our literature-consistency check.

Task	Model	Run	P Check	K Check	Main Mismatch
real_018	GPT-5.2	Run 1	all matched planets pass	fail	HD 69830 b/c/d recovered at $P \approx 8.666, 31.559, 197.330$ d, but with inflated semi-amplitudes: $K = 8.30/7.23/2.87\,\mathrm{m\,s^{-1}}$ versus literature $3.51/2.66/2.20\,\mathrm{m\,s^{-1}}$.
real_018	Claude-Sonnet-4.6	Run 2	all matched planets pass	fail	Same qualitative solution as GPT-5.2 on HD 69830, with all three periods matching the literature system but all three K values overestimated by several σ.
real_018	Gemini-3.1-Pro	Run 2	all matched planets pass	fail	Again recovers the HD 69830 three-planet period structure accurately, but returns $K = 8.30/7.23/2.87\,\mathrm{m\,s^{-1}}$ scale amplitudes rather than the published lower-amplitude solution.
real_020	Gemini-3.1-Pro	Run 2	pass	fail	HD 88133 b is recovered at the correct short period ($P \approx 3.416$ d), but the submitted semi-amplitude is $K = 77.6\,\mathrm{m\,s^{-1}}$ versus the literature value of $36.1\,\mathrm{m\,s^{-1}}$.
real_018	GPT-5.3-codex	Run 3	all matched planets pass	fail	Recovers the HD 69830 periods, but still overestimates all three amplitudes ($K = 6.65/7.82/3.42\,\mathrm{m\,s^{-1}}$) relative to the literature solution.

Table 6. Closest literature matches among agent submissions on real-data tasks. No submission across the three evaluation runs simultaneously matched all literature periods and semi-amplitudes within the reported uncertainties. The five entries shown here are the closest cases: all matched planets have literature-consistent periods, but at least one semi-amplitude K remains outside the published uncertainty interval.

B. Per-Criterion Analysis

B.1. Pass Rate Breakdown by Criterion

	Easy				Medium				Hard			
Model	ΔBIC	RMS	Match	Count	ΔBIC	RMS	Match	Count	ΔBIC	RMS	Match	Count
Kimi-K2.5	100	50.0	50.0	100	89.8	72.7	27.3	77.3	84.9	64.2	1.9	35.8
Qwen-3.5-Plus	90.0	85.0	80.0	100	87.5	87.5	36.4	80.7	89.8	71.2	0.0	27.1
o3-mini	98.3	80.8	70.0	82.5	92.5	78.6	27.7	69.2	89.7	62.3	0.7	27.4
GPT-5.2	100	93.1	86.2	100	94.0	96.4	43.4	77.1	95.2	90.5	11.9	59.5
Claude-Sonnet-4.6	97.9	97.9	87.2	95.7	100	98.7	46.1	82.9	96.6	96.6	3.4	10.3
Gemini-3.1-Pro	100	95.8	91.7	100	95.5	97.3	39.6	76.6	88.8	86.7	3.1	58.2
GPT-5-mini	100	92.7	83.6	92.7	93.3	94.3	39.0	73.3	93.3	84.4	0.0	25.6
GPT-5.3-codex	94.7	84.2	78.9	94.7	92.6	90.7	38.9	79.6	100	100	33.3	33.3

Table 7. Per-criterion pass rates (%) among tasks with $\geq$1 submission, broken down by difficulty tier. ΔBIC and RMS measure *statistical detection*; Match and Count measure *physical recovery*.

Table 7 breaks down the pass rate for each of the four evaluation criteria (ΔBIC, RMS, Match Score, Planet Count) by difficulty tier, revealing several patterns not visible in the aggregate pass rates.

Statistical criteria do not distinguish models. On Easy tasks, ΔBIC exceeds nearly 90% for almost all models and RMS remains uniformly high. Detecting a periodic signal and achieving a reasonable fit is not a bottleneck; the standard periodogram-to-Keplerian pipeline is well within frontier model capabilities.

Match Score and Planet Count capture distinct failure modes. On Hard tasks, Planet Count remains above 25% for several models (GPT-5.2: 59.5%, Gemini: 58.2%), indicating that some agents correctly infer the number of planets but recover inaccurate orbital parameters. Match Score collapses to single digits for most models (<5%), revealing that the bottleneck is not merely deciding *how many* planets exist, but accurately *characterising* their orbits.

Claude-Sonnet-4.6: best fits, worst planet count. Claude achieves the highest statistical rates across all tiers (Hard: 96.6% ΔBIC, 96.6% RMS), yet its Hard-tier Planet Count is the lowest of all models (10.3%). It excels at fitting within a fixed model but systematically fails at model selection.

Conjunction gate is strict. GPT-5.3-codex achieves 100% on both statistical criteria on Hard tasks, yet its overall Hard pass rate is only 4.2%, because Match and Count must *both* pass simultaneously.

Implication. The core bottleneck is *model selection* (choosing the correct number of planets) and *parameter recovery* (accurately determining orbital elements). Both require physical reasoning beyond optimisation: knowing when to escalate model complexity, recognising alias periods, and diagnosing whether structured residuals reflect missing planets or correlated noise.

Stricter match score aggregation. The current implementation averages S_{match} over successfully paired planets only; unmatched truth planets (with pairwise distance $d_{ij} > 5$) do not contribute to the score. A stricter formulation would normalize by $|T|$ (the number of true planets) rather than $|\mathcal{M}|$ (the number of matched pairs):

$$S_{\text{match}} = \frac{1}{|T|} \sum_{(i,j)\in\mathcal{M}} e^{-d_{ij}} - 0.25\,|n_{\text{truth}} - n_{\text{guess}}|\,. \tag{5}$$

Under this formulation, a submission that correctly identifies one of three planets cannot pass `ok_match` regardless of how well that planet is recovered. We retain the current mean-over-matched formulation as the primary metric because frontier models rarely trigger this edge case under current capability levels, but recommend adopting the stricter variant as models improve and partial-recovery solutions become more common.

B.2. Output Format Compliance

As noted in Section 4.2, Qwen-3.5-Plus and Kimi-K2.5 frequently produce malformed JSON when submitting analysis results. Our harness includes a regex-based fallback parser that attempts to repair common formatting errors (e.g., trailing commas, unquoted keys, truncated brackets). Despite this mitigation, a substantial fraction of malformed outputs remain unrecoverable, continuing to consume step budget without advancing the analysis. The reported pass rates for these two models therefore reflect performance *after* automated repair; without it, their scores would be even lower.

B.3. Difficulty Factor Correlations

Figure 5 reports the Pearson correlation between difficulty factors and per-task binary success (computed over all 100 synthetic tasks and 3 runs per model). Low SNR and higher planet multiplicity are the strongest negative predictors.

B.4. Per-Criterion Effect of Skills Injection

Table 8 decomposes the skills ablation (Section 4.3) into the four evaluation criteria for all four models. Three patterns emerge:

Easy tier: uniform procedural improvement. All four models show improved or stable ΔBIC, RMS, and Count on Easy tasks after skills injection. Gemini-3.1-Pro reaches 100% on all four criteria. This confirms that skills successfully encode the standard periodogram-to-fit workflow.

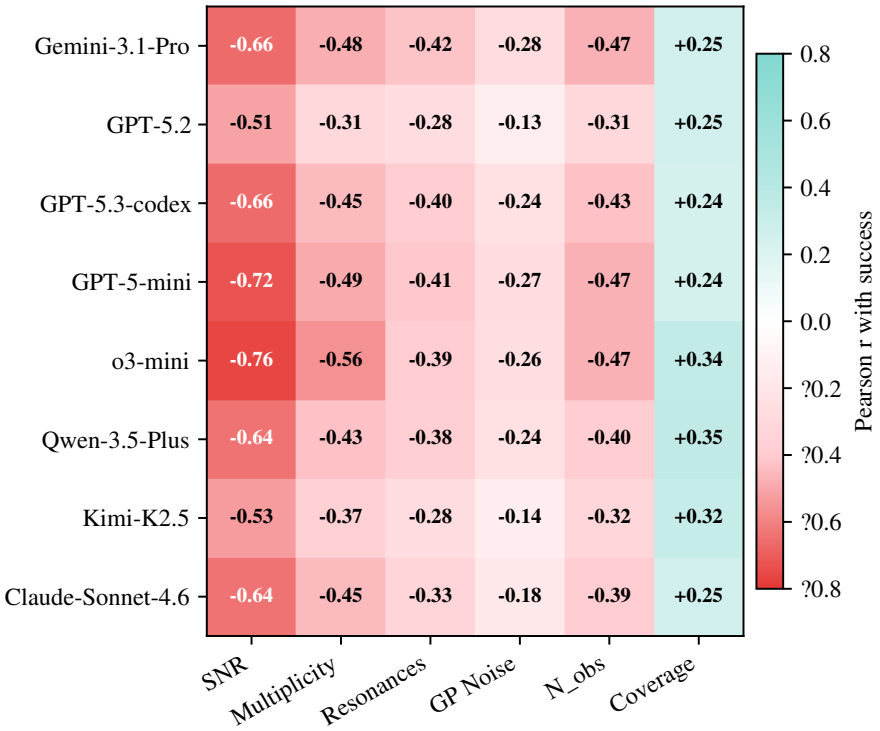

Figure 5. Pearson correlation between difficulty factors and per-task success, aggregated over 100 synthetic tasks and 3 runs per model.

	Easy				Medium				Hard			
Model	ΔBIC	RMS	Match	Count	ΔBIC	RMS	Match	Count	ΔBIC	RMS	Match	Count
GPT-5-mini	100.	$98.3_{+2.9}$	$90.0_{+1.9}$	$98.3_{+3.8}$	91.7.	$85.8_{-7.9}$	$38.3_{+2.6}$	$79.2_{+8.6}$	$94.9_{+3.4}$	$72.9_{-12.7}$	2.5.	$33.9_{+2.5}$
Qwen-3.5-Plus	$96.1_{+3.4}$	$92.2_{+9.2}$	$86.3_{+5.8}$	100.	$89.8_{+3.0}$	86.7.	$38.8_{+3.6}$	$81.6_{+2.5}$	$91.5_{+5.6}$	$66.0_{-7.3}$	$0.0_{-2.8}$	$25.5_{-4.0}$
Gemini-3.1-Pro	100.	$100_{+3.1}$	$100_{+7.2}$	100.	$90.4_{-5.1}$	$93.6_{-3.7}$	$81.9_{+42.6}$	$94.7_{+17.9}$	$88.7_{+2.6}$	$87.3_{+1.2}$	$32.4_{+26.5}$	$54.9_{-7.4}$
GPT-5.3-codex	96.2.	86.8.	$83.0_{-1.2}$	$98.1_{+3.4}$	$92.4_{-2.1}$	$84.8_{-6.3}$	$63.8_{+23.8}$	$98.1_{+21.4}$	$85.1_{-14.9}$	$75.9_{-10.8}$	$32.2_{-7.8}$	$85.1_{+31.7}$

Table 8. Per-criterion pass rates (%) *with* skills injection, averaged over three runs, among tasks with ≥1 submission. ΔBIC and RMS are statistical criteria; Match and Count are physical criteria. Superscripts show the change vs. the default condition in Table 7 (+ improvement, − degradation, · <1 pp).

Medium/Hard: strong models gain on physical criteria. The most striking effect is in Match Score on Medium tasks: Gemini-3.1-Pro jumps from 39.3% to 81.9% (+42.6 pp) and GPT-5.3-codex from 40.0% to 63.8% (+23.8 pp). Planet Count shows similar gains for these two models (+17.9 and +21.4 pp on Medium). This indicates that procedural scaffolding helps stronger models land closer to correct orbital solutions, not just achieve better statistical fits.

Weaker models: RMS degrades on Hard tasks. GPT-5-mini and Qwen-3.5-Plus show RMS degradation on Hard tasks (−12.7 and −7.3 pp respectively), while their Match Scores remain near zero. This suggests that the procedural template imported from Easy-tier trajectories may actively interfere with the more exploratory fitting strategies required for complex multi-planet systems.

B.5. Episode Termination Under Skills Injection

Table 9 decomposes each episode into two mutually exclusive outcomes: the agent finishes naturally (**Env Done**) or is cut off by a resource limit (**Budget Exceeded**).

The key observation is that skills shift episodes *from* Budget Exceeded *to* Env Done without changing what the agent does once it reaches the submission stage. For Gemini-3.1-Pro on Hard tasks, Budget Exceeded drops from 93.5% to 83.3%

(-10.1 pp), exactly matching the Env Done increase. This means the pass-rate gains reported in Table 2 are almost entirely attributable to *efficiency*: skills compress the workflow so that episodes that previously timed out now reach the submission stage. The underlying physical reasoning is not improved, as Table 8 confirms (Match Score on Hard tasks remains below 33% for all models).

		Easy		Medium		Hard	
Model	**Variant**	Pass	Submitted	Pass	Submitted	Pass	Submitted
GPT-5-mini	Default	76.7	76.7	31.7	48.3	2.5	5.8
	+Skills	$90.0_{+13.3}$	$90.0_{+13.3}$	$35.8_{+4.1}$	$49.2_{+0.9}$	$2.5_{+0.0}$	$5.0_{-0.8}$
Qwen-3.5-Plus	Default	26.7	26.7	25.0	25.8	1.7	2.5
	+Skills	$48.3_{+21.6}$	$48.3_{+21.6}$	$23.3_{-1.7}$	$23.3_{-2.5}$	$0.0_{-1.7}$	$0.0_{-2.5}$
Gemini-3.1-Pro	Default	71.7	71.7	34.2	35.0	5.0	5.8
	+Skills	$100_{+28.3}$	$100_{+28.3}$	$56.7_{+22.5}$	$58.3_{+23.3}$	$16.7_{+11.7}$	$16.7_{+10.9}$
GPT-5.3-codex	Default	80.0	88.3	29.2	49.2	4.2	7.5
	+Skills	$74.6_{-5.4}$	$75.0_{-13.3}$	$53.2_{+22.4}$	$71.7_{+22.5}$	$25.6_{+21.4}$	$48.3_{+40.8}$

Table 9. Skills injection: Pass Rate vs. Submission Rate (%), averaged over three runs. **Pass** = all four criteria satisfied; **Submitted** = agent produced at least one submission before budget exhaustion. Superscripts: green = improvement, red = degradation vs. Default.

B.6. Match Score Threshold Sensitivity

Table 10 reports pass rates under $\pm 10\%$ variation of the Match Score threshold (default 0.80). The match-score distribution is strongly bimodal: most submissions score either >0.9 (correct recovery) or <0.5 (wrong parameters), with only 17% of submissions falling in the 0.70–0.90 boundary region. As a result, threshold variation produces <5 pp change in overall pass rates and preserves relative model rankings across all tiers.

Table 10. Sensitivity of pass rate (%) to the Match Score threshold. The default threshold is 0.80; columns show results under $\pm 10\%$ variation. Top-tier rankings on Easy and Medium tasks are largely preserved; reorderings on Hard tasks reflect near-zero absolute rates where single-task differences dominate.

	$S_{\text{match}} \geq 0.72$ (-10%)			$S_{\text{match}} \geq 0.80$ (default)			$S_{\text{match}} \geq 0.88$ ($+10\%$)		
Model	Easy	Med	Hard	Easy	Med	Hard	Easy	Med	Hard
o3-mini	40.5	27.2	0.5	40.0	24.2	0.0	37.9	19.9	0.0
GPT-5-mini	81.7	39.9	3.7	76.7	33.6	2.5	76.7	28.2	0.0
GPT-5.2	45.0	39.0	5.8	40.0	30.0	5.8	40.0	29.9	0.0
Kimi-K2.5	15.0	20.7	0.8	13.3	17.9	0.8	13.3	15.1	0.8
Qwen-3.5-Plus	28.4	31.4	1.6	26.7	25.0	1.6	26.7	22.2	0.0
Gemini-3.1-Pro	76.7	40.9	8.1	71.7	35.0	5.0	70.0	29.9	0.0
Claude-Sonnet-4.6	73.4	26.4	0.8	68.3	22.5	0.8	68.3	16.7	0.0
GPT-5.3-codex	81.7	38.0	4.2	80.0	30.8	4.2	78.3	24.6	0.0

C. Domain-Expert Skills

In our skills-injection experiments (§4.3), each agent receives five domain-expert skills appended to its system prompt. Below we reproduce the full text of each skill exactly as provided to the agent.

Skill 1: RV Period Search & Alias Detection

Description. Use this skill when searching for planetary orbital periods in RV data, especially to avoid picking up 1-day aliases or harmonics instead of the true period.
When to Activate.

- Running a periodogram (GLS, Lomb-Scargle) on RV data
- Choosing between multiple peaks in a periodogram
- Unsure whether a period candidate is real or an alias
- After fitting a planet and residual RMS is unexpectedly high

Instructions.

1. **Always detrend first.** Remove linear or polynomial RV drift before running any periodogram. Failure to detrend causes spurious long-period peaks.
2. **Reject periods > baseline.** If any periodogram peak has period > (max_time − min_time), it is almost certainly a trend alias. Reject immediately.
3. **Identify the 1-day alias family.** For candidate P, compute: $1/(1/P - 1)$, $1/(1/P + 1)$, $P/2$, $2P$. If a secondary peak lies at one of these, pick the one with higher power and shorter period.
4. **Validate with phase-folding.** Phase-fold at top-3 candidates. Random scatter = wrong period; smooth curve = correct period.
5. **Narrow refinement.** Refine with fine grid search around $\pm 5\%$ of best candidate.

Skill 2: Robust Keplerian Orbit Fitting

Description. Fit a full Keplerian orbit model to RV data. Ensures eccentricity is properly optimized, avoids local minima, and produces reliable parameter estimates.
When to Activate.

- After identifying a candidate period from a periodogram
- When a sine-fit gives poor RMS or implausible eccentricity
- When fitting 1 or more planets to RV data

Instructions.

1. **Never use sine-fit as final answer.** Sine fitting assumes $e = 0$. Always follow up with a full Keplerian fit with 6 parameters: $P, K, e, \omega, M_0, \gamma$.
2. **Use global optimization.** Local optimizers get stuck in local minima. Use `differential_evolution` with bounds: $P \pm 2\%$ of periodogram peak, $K \in [0.1, 3K_{\text{sine}}]$, $e \in [0, 0.8]$, ω, $M_0 \in [0, 2\pi]$.
3. **Polish with local optimizer.** After global search, refine with Nelder-Mead (`maxiter=10000`).
4. **Validate eccentricity.** If $e > 0.8$: suspect artifact, re-fit with $e \leq 0.7$. If RMS with $e = 0$ is within 5% of best-fit RMS: submit circular orbit.
5. **Per-instrument offsets.** If multiple instruments, fit independent γ_i for each.
6. **Report RMS.** If RMS $\gg \sigma_{\text{median}}$, try different period or starting eccentricity.

Common Mistakes. Do not submit a sine-fit directly. Do not fix $e = 0$ unless justified. Do not use only local optimization. Do not use $t = 0$ as reference—always use $t_{\text{ref}} =$ `times[0]`.

Skill 3: Mean Longitude (l_{rad}) Calculation

Description. Correctly compute the mean longitude l_{rad} at the reference epoch $t_{\text{ref}} =$ `times[0]`, required for `submit_action`. This is the most common source of `ok_match` failure.
The correct formula:

```
l_rad = (Omega_rad + omega_rad
         + M0_at_tref) mod 2*pi
```

For RV-only fits (no astrometry), $\Omega = 0$, so:

```
l_rad = (omega_rad + M0_at_tref) % (2*pi)
```

Computing M_0 at t_{ref}. If your optimizer gives M_0 at some time t_{fit}, convert:

```
M0_at_tref = (M0_at_tfit
```

```
            + 2*pi/P * (t_ref - t_fit)) % (2*pi)
```

If M_0 was defined directly at `times[0]` during fitting (recommended), then $M_0 = M_{0,\mathrm{tref}}$ directly.
Sanity checks. (1) $l_{\mathrm{rad}} \in [0, 2\pi]$. (2) Verify RV at t_{ref} matches the first observed RV. (3) Compute l_{rad} independently for each planet.
Critical: Always use `times[0]` as reference epoch. Using $t = 0$, JD 2450000, or the midpoint will cause `ok_match` failure even if all other parameters are correct.

Skill 4: Multi-Planet Detection via Iterative Residual Analysis

Description. Determine the correct number of planets by iteratively fitting and subtracting signals. Multi-planet tasks are the most common cause of `ok_count` failure.
When to Activate.

- After fitting a first planet and computing residuals
- When task difficulty tier is medium to hard.
- When residual RMS after first planet fit is still $> 2\times$ the noise floor

Instructions.

1. **Check residuals.** Compute residual RMS after subtracting planet 1. If RMS $> 2 \times \sigma_{\mathrm{median}}$, a second planet likely exists. If RMS $\leq 1.5 \times \sigma_{\mathrm{median}}$, likely single-planet.
2. **Periodogram on residuals.** Run Lomb-Scargle on residuals, apply alias checks from Skill 1.
3. **BIC comparison.** Compute BIC for N-planet vs $(N+1)$-planet model. $\Delta\mathrm{BIC} > 10$: strong evidence; > 6: moderate evidence; < 6: no strong evidence, stop.
4. **Joint re-optimization.** After finding approximate P_2 from residuals, fit both planets simultaneously.
5. **Repeat.** Check 2-planet residuals for a 3rd planet. Stop when residual RMS $\approx$ noise floor or $\Delta\mathrm{BIC} < 6$.

Decision Rule: Residual RMS $> 3\times$ noise $\rightarrow$ add planet; 2–3$\times$ noise $\rightarrow$ check ΔBIC; $< 2\times$ noise $\rightarrow$ stop.

Skill 5: Submission Strategy & Timing

Description. Decide when to call `submit_action` and avoid analysis paralysis (running out of budget without submitting). No-submission is the worst outcome (reward $= 0$).
The Golden Rule: Submit Early, Refine Later. Submit a baseline solution as soon as you have a fit with RMS below a reasonable threshold. You can always submit again—only the best submission counts.
Submit if ANY of these are true:

- RMS < 30 m/s (for typical tasks)
- You have spent > 8 tool calls without submitting
- Your current fit matches the period and K within 10%
- You are about to try something risky (MCMC, re-fit from scratch)

Time Budget Allocation.

- *Phase 1—Exploration* (calls 1–8): Load data, periodogram, rough sine fit.
- *Phase 2—First Fit* (calls 9–16): Full Keplerian, compute l_{rad}. $\rightarrow$ **Submit baseline here.**
- *Phase 3—Refinement* (calls 17–25): Check residuals for additional planets, refine. $\rightarrow$ Submit improved solution if better.
- *Phase 4—Polish* (calls 26+): MCMC only if time permits. Do not start from scratch.

Emergency Protocol. If close to budget with no submission: immediately fit a circular Keplerian ($e = 0$) using the best periodogram period and submit. A rough solution is infinitely better than no submission.

D. System Prompt

Below is the complete system prompt sent to each agent at the start of a task, shown for a representative Easy-tier task (seed 44, difficulty 2). Dynamic fields (marked with boxes) are populated at runtime from each task's observation data; all other text is shared across tasks. The prompt specifies the agent's role, available tools, submission format, a mandatory six-step analysis strategy, and common pitfalls. Budget constraints and fit-quality thresholds scale with task difficulty.

System Prompt *Example: seed 44 | Easy | 1 planet*

You are an expert RV data analyst tasked with detecting exoplanets from radial velocity measurements.

Dataset Overview

- Number of observations: [80] - Time span: [33.5] days - RV range: [-88.10] to [+74.56] m/s - Median uncertainty: [2.60] m/s (boxed = dynamic)

Your Task

Analyze the RV data to identify planetary signals. Use the PythonREPL tool to:
- Compute periodograms (Lomb-Scargle or other methods) - Test baseline models (via baselines module)
- Fit Keplerian orbital models (NOT simple sinusoids) - Analyze residuals and trigger optimization when needed

CRITICAL: Keplerian Model Parameters

When fitting Keplerian orbits, you MUST fit ALL of these parameters:
- **P**: Period (days) - **K**: RV semi-amplitude (m/s) - **e**: Eccentricity (0--0.8)
- **omega**: Argument of periastron (radians, 0 to 2pi) --- CRITICAL FOR ECCENTRIC ORBITS!
- **M0**: Mean anomaly at reference time (radians) - **gamma**: Systemic velocity offset (m/s)

Submission Format

Use Stargazer-native planet fields: P_days, m_sin_i_mjup, e, omega_rad, l_rad
l_rad = mean longitude at reference epoch t_ref = times_days[0]. Convert: l_rad = (Omega_rad + omega_rad + M0) % (2*pi)

Response format for every turn

1) Findings: concise hypothesis plus key numbers. 2) Plan/Next: 1--3 short bullets.
3) Code: one fenced code block (only if calling PythonREPL). 4) Results: outputs interpreted; if ready, submit.

Available Tools

1. **PythonREPL**: Execute Python code. Pre-loaded: times_days, rvs_ms, sigmas_ms, np, baselines, history, star_mass_sun, t_ref_days, Stargazer_planet_from_fit, Stargazer_SUBMISSION_GUIDE
2. **submit_action**: Submit planet hypotheses (max [3] planets, period > 0.5d, eccentricity 0--0.8)

Budget Constraints

- Max tool calls: [12] - Max execution time: [300.0]s - Max submissions: [3] (scaled by difficulty)

Mandatory Step 0: Read Protocol Guide First

Before any fitting/submission, read Stargazer_SUBMISSION_GUIDE in PythonREPL and set: _protocol_guide_ack = True.

Strategy (FOLLOW THIS ORDER)

Step 1: Periodogram Analysis --- Compute Lomb-Scargle periodogram; identify strongest peak(s).
Step 2: Linear Sine Baseline (MANDATORY) --- Run baselines.baseline_one_sine(observation); print period, RMS, RMS/sigma.
Step 3: Model Gating Decision --- If RMS/median_sigma <= 1.5: submit directly; else: escalate to Kepler.
Step 4: Keplerian Fitting (ONLY IF GATE=YES) --- Full 6-param fit; multi-start optimisation for e > 0.3.
Step 5: Check Fit Quality --- Good fit: RMS approx [2.60] m/s; bad fit: keep optimising.
Step 6: Submit ONLY After Convergence --- Include ALL parameters, especially omega_rad!

Common Mistakes to AVOID

1. Jumping to Kepler before LS + linear-sine gating 2. Submitting early with poor fit (high RMS)
3. Forgetting omega_rad in submission (defaults to 0!) 4. Wrong phase convention for l_rad
5. Not doing multi-start optimisation for eccentric orbits

E. Case Study Trajectories

Figure 6 shows the RV fits for two representative case studies.

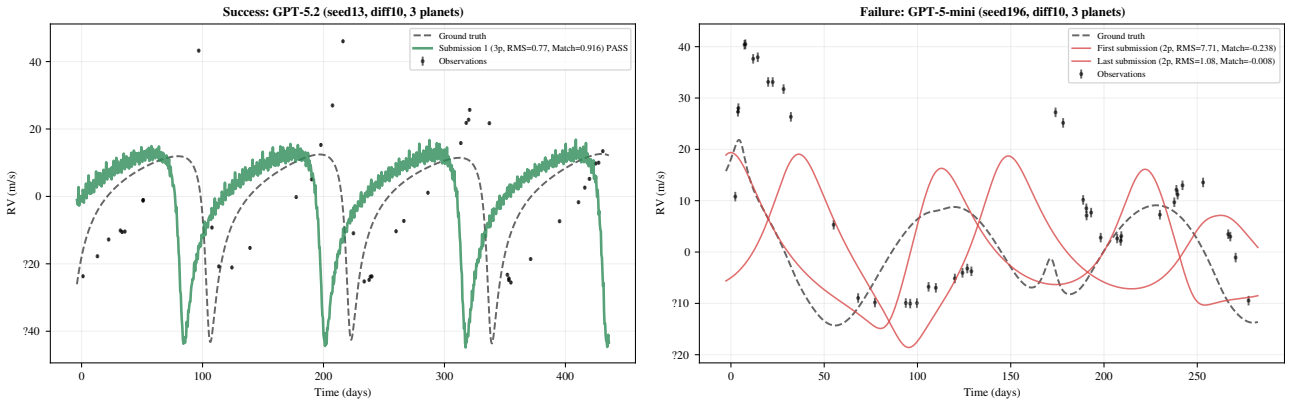

Figure 6. RV model fits for two case studies. **(a)** GPT-5.2 recovers both planets on seed96 (Medium); the agent's fit (green dashed) closely matches the ground truth (blue solid). **(b)** GPT-5-mini fails on seed196 (Hard): its best 2-planet submission converges to alias periods ($P = 111.7,\ 164.5$ d instead of the true $76,\ 112,\ 167$ d), producing an RV curve (red dashed) that visibly diverges from the three-planet ground truth.

✓ Success Case: 2-Planet Recovery (Medium)

GPT-5.2 | 11 steps

Task Seed 96

Planets	2 ($P = 31.2,\ 117.6$ d)	**Observations**	75
Noise model	White ($\sigma_w = 1.05\,\mathrm{m\,s^{-1}}$)	**SNR**	8.9

Difficulty. Two-planet system with moderate SNR. Planet 2 dominates the RV signal ($K_2 \approx 30\,\mathrm{m\,s^{-1}}$, $e_2 = 0.25$), but its period (117.6 d) exceeds the data span (106.7 d), causing the initial periodogram to converge on a biased estimate (~ 105 d). Planet 1 ($P = 31.2$ d, $K_1 \approx 14\,\mathrm{m\,s^{-1}}$) is buried in the residuals of the dominant signal.

Steps 1–5 Protocol & Baseline

Reads protocol, computes weighted LS periodogram (40 000 trial frequencies). Dominant peak near $P \approx 105.0$ d (power $= 0.800$). Baseline one-sine fit: $P = 104.8$ d, $K = 20.3\,\mathrm{m\,s^{-1}}$, RMS $= 10.68\,\mathrm{m\,s^{-1}}$ (RMS/$\tilde{\sigma} = 10.1$). Gate decision: **Kepler = YES.**

Steps 6–9 Iterative Keplerian Fitting

1-planet fit (multi-start): converges to $P = 300$ d at the search boundary—a clear sign the true period exceeds the data span. RMS $= 9.16\,\mathrm{m\,s^{-1}}$. Residual periodogram reveals a strong second signal at $P \approx 31.0$ d (power $= 0.842$).
2-planet fit (multi-start): $P_1 = 116.6$ d, $K_1 = 29.3\,\mathrm{m\,s^{-1}}$, $e_1 = 0.25$; $P_2 = 31.1$ d, $K_2 = 14.0\,\mathrm{m\,s^{-1}}$, $e_2 = 0.02$. RMS: $9.16 \rightarrow \mathbf{1.13}\,\mathrm{m\,s^{-1}}$ (RMS/$\tilde{\sigma} = 1.07$).

Step 11 Submit #1 (2 planets) → Pass

Match $= 0.932$, RMS $= 1.13\,\mathrm{m\,s^{-1}}$, ΔBIC/pt $= 513$. ✓bic ✓rms ✓match ✓count

Resources: 61 K input + 7 K output tokens, 0 code errors, 1 submission.

Takeaway. The agent follows a textbook *peel-and-search* strategy: fit the dominant signal, subtract, search residuals for the next. A key diagnostic moment occurs at Step 7: the 1-planet fit converges to the period upper bound ($P = 300$ d), prompting the agent to recognise incomplete convergence and check residuals rather than submit. The strong residual peak at 31 d guides the 2-planet fit, which simultaneously refines P_1 from the boundary value to the correct 116.6 d. Both planets are recovered in 11 steps with a single submission, consuming only 68 K tokens—the most efficient multi-planet recovery in the benchmark.

✗ Failure Case: Stuck in a Local Minimum (Hard)

GPT-5-mini | 40 steps

Task Seed `196`

Planets	3 ($P = 76,\ 112,\ 167$ d)	**Observations**	43
Noise model	White ($\sigma_w = 0.93\,\mathrm{m\,s^{-1}}$)	**Resonance**	Double ($\approx$ 3:2:1)

Difficulty. Three planets in a double-resonance chain ($\approx$ 3:2:1), producing alias peaks and combination frequencies. The strongest periodogram peak ($\sim$ 117 d) is shifted from the true $P_2 = 112$ d by unmodelled planets at 76 and 167 d.

Steps 1–12 Protocol, Baseline & Fitting

LS periodogram finds $P \approx 117.9$ d. Baseline: $P = 116.9$ d, RMS $= 3.35\,\mathrm{m\,s^{-1}}$ (RMS$/\tilde{\sigma} = 3.6$). Gate: **Kepler = YES.** 1-planet fit (108 multi-starts): $P_1 = 116.11$ d, $e_1 = 0.183$. RMS $= 3.06\,\mathrm{m\,s^{-1}}$. Residual periodogram reveals $P_2 \approx 165.9$ d. 2-planet fit: $P_1 = 111.73$ d, $P_2 = 164.52$ d. RMS $= \mathbf{1.08}\,\mathrm{m\,s^{-1}}$.
Neither detected period matches a true planet — both are alias frequencies.

Steps 13–16 Submissions #1–#2 → Format Errors

Submit #1: `l_rad` set to 0 ⇒ RMS $= 7.71\,\mathrm{m\,s^{-1}}$. Reward $= -24.6$. Submit #2: wrong `inc_rad` ⇒ RMS $= 9.39\,\mathrm{m\,s^{-1}}$. Reward $= -45.3$. *Two submissions wasted on format debugging before the model itself is evaluated.*

Steps 17–20 Submit #3 (correct format) → Alias Fail

Properly formatted submission evaluates at RMS $= 1.08\,\mathrm{m\,s^{-1}}$ (passes `ok_rms`), but **match** $= -0.008$: the detected periods $(111.7, 164.5)$ d are alias periods of the true near-resonant system. Reward $= 18.2$.

Steps 21–40 Submissions #4–#10 → Perseveration

Seven further submissions: the agent resubmits the same 2-planet alias solution 4 times identically (reward $= 18.2$), tries a 1-planet fallback (reward $= 11.9$), and attempts minor parameter variations. **Never attempts a 3-planet model** despite clear residual structure. No new period candidates are explored in the final 20 steps.

Resources: 696 K input + 34 K output tokens, 14 code errors, 10 submissions — all failed.
Diagnostics: ✓bic ✓rms ✗match ✗count (2 vs. 3 true)

Takeaway. Three compounding failure modes: *(1) Format fragility*: 30% of submission budget wasted on API errors (`l_rad`, `inc_rad`). *(2) Alias convergence*: the detected periods $(111.7, 164.5)$ d are combination frequencies of the true 3:2:1 resonant system — a classic RV pitfall where aliased periods produce low RMS but match ≈ 0. *(3) Perseveration*: the same wrong answer is resubmitted 4 times without model escalation. The failure consumes 10.7× more tokens (730 K vs. 68 K) than the success case yet produces a negative match score, demonstrating that more computation does not compensate for the inability to revise the model hypothesis.