Plausible but Wrong: A Case Study on Agentic Failures in Astrophysical Workflows

Shivam Rawat [1] **Lucie Flek** [1]

Abstract

Agentic AI systems are increasingly being integrated into scientific workflows, yet their behavior under realistic conditions remains insufficiently understood. We evaluate CMBAgent across two workflow paradigms and eighteen astrophysical tasks. In the One-Shot setting, access to domain-specific context yields an approximately $\sim 6\times$ performance improvement (0.85 vs. ≈ 0 without context), with the primary failure mode being silent incorrect computation—syntactically valid code that produces plausible but inaccurate results. In the Deep Research setting, the system frequently exhibits silent failures across stress tests, producing physically inconsistent posteriors without self-diagnosis. Overall, performance is strong on well-specified tasks but degrades on problems designed to probe reasoning limits, often without visible error signals. These findings highlight that the most concerning failure mode in agentic scientific workflows is not overt failure, but confident generation of incorrect results. We release our evaluation framework to facilitate systematic reliability analysis of scientific AI agents.

1. Introduction

Recent advances in large language models (LLMs) (Brown et al., 2020; Wei et al., 2022; Achiam et al., 2023) and tool-augmented reasoning systems (Yao et al., 2022; Schick et al., 2023; Nakano et al., 2021) have enabled the emergence of agentic AI systems capable of multi-step reasoning, external tool use, and autonomous decision-making. These systems are increasingly proposed as assistants for scientific workflows, where they can perform data analysis, hypothesis generation, and pipeline orchestration with limited human intervention.

[1]Bonn-Aachen International Center for Information Technology, University of Bonn, Germany; Lamarr Institute for Machine Learning and Artificial Intelligence, Germany. Correspondence to: Shivam Rawat <s.rawat@uni-bonn.de>.

Preprint. June 2, 2026.

Despite promising demonstrations, existing work largely evaluates such systems through task-level success or synthetic benchmarks, offering limited insight into their reliability in realistic scientific settings. Scientific workflows differ fundamentally from standard AI benchmarks: they involve structured pipelines, domain-specific constraints, noisy data, and multi-stage reasoning where early errors can propagate throughout the pipeline. As a result, aggregate success metrics alone are insufficient to characterise system behaviour in these environments.

Importantly, our focus is not on proposing or critiquing agentic architectures themselves, but on evaluating their scientific reliability and failure modes when deployed in realistic workflows.

A central challenge is that agentic systems inherit known limitations of LLMs, including hallucination, brittle reasoning, and misalignment with task objectives. When embedded in multi-step workflows with external tool interaction, these issues can lead not only to incorrect intermediate steps but also to silent numerical errors and physically inconsistent outputs. Critically, such failures may not be detectable from final outputs alone, making systematic evaluation of reliability and error modes essential for scientific deployment.

In this work, we present a structured empirical evaluation of an existing agentic system applied to astrophysical inference tasks. Rather than introducing a new model or framework, our goal is to characterise performance boundaries, assess robustness and calibration, and identify systematic failure modes in realistic scientific workflows.

Astrophysics provides a particularly suitable testbed due to its well-defined physical models, established numerical tools, and availability of reference solutions from both simulation and observation. This allows controlled evaluation of correctness while maintaining real-world scientific relevance.

We introduce a structured evaluation framework for agentic scientific workflows that integrates execution success, parameter accuracy, and numerical fidelity. Using this framework, we conduct a systematic analysis of reliability across both single-step and multi-step reasoning settings, iden-

tifying key failure modes such as silent numerical errors and physically inconsistent inference, and quantifying their prevalence under different task conditions.

2. Related Work

LLM Agents and Tool Use. The ability of LLMs to interact with external tools has been a key driver in the transition from static text generation to dynamic agentic behaviour. ReAct (Yao et al., 2022) established the paradigm of interleaving reasoning traces with action execution, enabling models to plan and adapt based on environmental feedback. Toolformer (Schick et al., 2023) extended this by training models to invoke APIs autonomously through self-supervision, while WebGPT (Nakano et al., 2021) demonstrated that grounding language models in live retrieval substantially improves factual reliability. More recent work has formalised tool-use evaluation, with benchmarks such as ToolBench (Qin et al., 2025) and TRAJECT-Bench (He et al., 2025) revealing systematic failure modes including similar-tool confusion, parameter-blind selection, and performance degradation with trajectory length. Together these works establish that structured tool access, rather than parametric knowledge alone, is what enables LLMs to operate effectively in multi-step workflows — a finding our results reinforce in the context of scientific computation.

LLM Agents in Scientific Workflows. A growing number of works demonstrate that agentic systems can execute substantial portions of real scientific workflows, spanning hypothesis generation, data analysis, and manuscript production (Ding et al., 2025; Sun et al., 2025; Villaescusa-Navarro et al., 2025; Moreno et al., 2026; Hellert et al., 2025). While these works establish the feasibility of agentic scientific automation, evaluation is typically limited to end-to-end task success or qualitative expert assessment. Systematic analysis of reliability, failure modes, and error propagation in multi-step scientific pipelines remains largely unexplored.

Reliability and Evaluation of LLMs. A substantial body of work examines the limitations of LLMs, including hallucination (Huang et al., 2025; Alansari & Luqman, 2025), factual inconsistency (Lin et al., 2022), and misalignment with user intent (Bai et al., 2022). Critically, recent surveys show that hallucinations in agentic settings differ qualitatively from single-step LLM failures: they emerge through multi-step sequential interactions involving tool use, memory, and inter-agent communication, making them substantially harder to detect from final outputs alone (Lin et al., 2025). More recent work has begun to formalise agent evaluation, distinguishing between tool-use capability, trajectory-level reasoning, and domain-specific performance (Yehudai et al., 2025; Chowa et al., 2026). However, systematic evaluation of reliability under realistic multi-step scientific workflows remains limited.

Positioning of This Work. In contrast to prior work, we focus explicitly on reliability and failure characterisation rather than task feasibility. Rather than proposing a new model or architecture, we provide a structured empirical evaluation of an existing deployed system under controlled conditions such as under-constrained inference and compositional reasoning pipelines. This enables us to identify not only whether agentic systems succeed, but how and why they fail, and crucially, whether those failures are visible in the final output or silent.

3. Agentic System Under Consideration

For our evaluation, we utilize *CMBAgent*, an open-source agentic framework for autonomous scientific discovery (Xu et al., 2025). *CMBAgent* serves as the research backend in the *Denario* project (Villaescusa-Navarro et al., 2025), generating research ideas, methodologies, and results through coordinated multi-agent planning. The outputs—including analyses, visualizations, and literature references—are automatically compiled into publication-ready manuscripts. While the *Denario* project demonstrates the feasibility of end-to-end autonomous scientific workflows and evaluates overall task success through expert assessment, a systematic investigation of reliability, robustness, error propagation, and failure modes remains limited. Such analysis is essential for assessing the suitability of agentic systems in scientific deployment settings where reproducibility and numerical correctness are critical. Here we complement existing feasibility studies by conducting a structured empirical evaluation focused on robustness, calibration, and failure characterization, thereby contributing toward a more rigorous evaluation framework for autonomous scientific agents.

The framework provides multiple operational modes and workflows, which differ primarily in the number of active agents, session length, context persistence across tasks, and the extent of human intervention required for conducting scientific analyses. These are as follows:

1. **One Shot.** Executes a task in a single reasoning and execution pass without iterative planning or review. We instantiate two variants: one with a context-aware domain agent (*cambagent*) providing task-relevant grounding, and one without, isolating the contribution of domain-specific retrieval augmentation relative to a base LLM baseline.

2. **Deep Research.** Activates the full Planning & Control architecture, decomposing tasks into substeps executed by specialised agents with critique and retry modules. Designed for complex, multi-step analyses requiring

multilayered reasoning and iterative error correction.

We focus exclusively on these two fully automated workflows; the **Human in the Loop** mode is excluded as user feedback introduces variability that complicates reproducible benchmarking (Yehudai et al., 2025; Chowa et al., 2026).

4. Evaluation and Metrics

We evaluate the agentic system across two distinct categories of tasks, aligned with the two workflow paradigms under investigation. This separation is intentional and reflects the differing operational objectives of the workflows. Tool-driven computational tasks are employed to evaluate the *One-Shot* workflow, as they emphasize single-pass reasoning, precise tool invocation, and accurate parameter configuration. These tasks allow controlled, fine-grained assessment of execution correctness and numerical reliability without introducing the additional complexity of multi-stage planning.

In contrast, the Deep Research workflow is specifically designed to support iterative reasoning, hierarchical task decomposition, and sustained context retention across multiple steps. Evaluating this workflow using simple, single-call computational tasks would not meaningfully exercise its planning and control mechanisms. Therefore, we assign more complex, research-oriented tasks to the Deep Research workflow—tasks that require multi-step reasoning, contextual carryover, and strategy refinement—so that its architectural capabilities can be appropriately tested. This task–workflow alignment ensures that each paradigm is evaluated under conditions that reflect its intended operational design.

Our evaluation design follows recent taxonomies of LLM-agents (Yehudai et al., 2025), which distinguish between capability-centric evaluation (e.g., tool use and function correctness), trajectory-level evaluation (e.g., planning and multi-step reasoning), and domain-specific application evaluation. The CAMB-based tasks primarily assess tool invocation accuracy and computational reliability, whereas the research-driven tasks stress hierarchical planning, long-horizon reasoning, and statistical inference consistency in scientific workflows.

4.1. Task Descriptions

Tool-Grounded Precision Tasks. Fourteen structured CAMB computation tasks assess parameter configuration robustness, solver reliability, and numerical accuracy, adapted from the CMBAgent benchmark repository (Contributors, 2024). We introduce a complexity stratification: Tasks 1–6 use a single API call; Task 7 adds tensor handling; Task 8 targets an alternate CAMB module; Tasks 9–10 require multi-API calls with combined results; Tasks 11–12 add a delensing pipeline; Task 13 adds ratio computation; and Task 14 combines multi-API calls with a noise-informed delensing pipeline.

Astrophysical Research-Driven Tasks. Four research-grade inference problems spanning cosmology, galactic dynamics, exoplanet structure, and strong gravitational lensing probe multi-step reasoning, statistical robustness, and resistance to hallucinated physical assumptions via Bayesian parameter estimation and hierarchical modeling. Task descriptions and category rationales are given in Section 4.3; prompts are in the Appendix.

4.2. Evaluation Metrics

We evaluate each workflow under a distinct scoring framework reflecting its operational objectives. The One-Shot workflow admits fully automated quantitative evaluation against reference outputs; the Deep Research workflow requires qualitative assessment against published literature values. All metrics are scored in $[0, 1]$ and averaged across trials for reporting.

4.2.1. One-Shot Workflow Metrics

Execution Success Rate (ESR). ESR is a binary indicator of whether the agent produced valid executable output. A trial is assigned $\mathrm{ESR} = 1$ if and only if the agent generates a numerical output file that (i) contains at least two numeric columns, (ii) covers at least 95% of the reference x-range, and (iii) provides at least 95% of the reference number of output points. Otherwise $\mathrm{ESR} = 0$. This criterion deliberately does not assess numerical correctness—it only verifies that the pipeline completed and produced an output of the right shape.

Parameter Accuracy Score (PAS). PAS measures whether the agent configured the CAMB cosmological solver with the correct input parameters. Parameters are extracted directly from the generated code via abstract syntax tree (AST) parsing, without executing the code, making the metric robust to runtime failures. For each parameter p present in the reference solution, the relative error is

$$\epsilon_p = \min\left(\frac{|\hat{\theta}_p - \theta_p^*|}{|\theta_p^*| + \delta},\ 1\right), \tag{1}$$

where θ_p^* is the reference value, $\hat{\theta}_p$ is the agent's value, and $\delta = 10^{-12}$ prevents division by zero. If the agent omits a required parameter, $\epsilon_p = 1$. The overall PAS is a weighted mean over all reference parameters:

$$\mathrm{PAS} = 1 - \frac{\sum_p w_p\, \epsilon_p}{\sum_p w_p}, \tag{2}$$

where weights w_p reflect the physical importance of each parameter: $w = 2.0$ for $\{H_0, \Omega_b h^2, \Omega_c h^2, n_s, A_s\}$, $w = 1.5$ for $\{\Omega_k, w_0\}$, and $w = 1.0$ for $\{\tau, m_\nu, A_{\rm lens}\}$. Weights reflect the breadth of spectral impact: core ΛCDM parameters (H_0, $\omega_b h^2$, $\omega_c h^2$, n_s, A_s; $w = 2.0$) distort amplitude, shape, and peak positions globally; extension parameters (Ω_k, w_0; $w = 1.5$) are task-critical only when non-zero; and secondary parameters (τ, m_ν, $A_{\rm lens}$; $w = 1.0$) have localised effects. This scheme underpenalises extension-parameter errors on tasks where they are non-trivial, making reported PAS a slight overestimate in those cases.

Numerical Accuracy Score (NAS). NAS measures how closely the agent's output curve matches the reference solution over the full output domain. The agent's curve is interpolated onto the reference grid and three complementary sub-metrics are computed.

NRMSE score measures amplitude accuracy:

$$S_{\rm NRMSE} = \max\left(0,\ 1 - \frac{\sqrt{\frac{1}{N}\sum_i (\hat{y}_i - y_i^*)^2}}{\max(y^*) - \min(y^*)}\right). \quad (3)$$

SMAPE score measures symmetric percentage deviation, robust to scale differences:

$$S_{\rm SMAPE} = \max\left(0,\ 1 - \frac{1}{N}\sum_i \frac{|\hat{y}_i - y_i^*|}{(|\hat{y}_i| + |y_i^*|)/2}\right). \quad (4)$$

CCC score is Lin's concordance correlation coefficient (Lin, 1989), which jointly penalises location shift, scale shift, and correlation:

$$S_{\rm CCC} = \max\left(0,\ \frac{2\,{\rm Cov}(\hat{y}, y^*)}{{\rm Var}(\hat{y}) + {\rm Var}(y^*) + (\bar{\hat{y}} - \bar{y}^*)^2}\right). \quad (5)$$

The sub-scores are combined as:

$$\text{NAS} = 0.2\, S_{\rm NRMSE} + 0.3\, S_{\rm SMAPE} + 0.5\, S_{\rm CCC}. \quad (6)$$

CCC receives the highest weight because it simultaneously captures amplitude, scale, and shape agreement; *NRMSE* is down-weighted because it is sensitive to outliers and can be dominated by a single high-amplitude peak.

Final Score.

$$\text{Score} = \begin{cases} 0 & \text{if ESR} = 0, \\ \text{PAS} \times \text{NAS} & \text{otherwise.} \end{cases} \quad (7)$$

The multiplicative form ensures that a trial scores zero if either parameters or numerical output are entirely wrong, even if the other dimension is perfect.

Failure Mode Taxonomy. Each trial is assigned to one of four mutually exclusive failure modes:

- **A Code failure** (ESR = 0): the pipeline did not produce valid output.
- **B Wrong parameters** (ESR = 1, PAS < 0.5): the code ran but used incorrect CAMB configuration.
- **C Wrong computation** (ESR = 1, PAS ≥ 0.5, NAS < 0.5): parameters were correct but numerical output was inaccurate, indicating a formula, unit, or post-processing error.
- **D Correct** (ESR = 1, PAS ≥ 0.5, NAS ≥ 0.5): both configuration and numerical output are consistent with the reference.

We additionally flag *unit/normalisation errors* as a sub-class of Mode C: trials where $S_{\rm CCC} > 0.8$ but $S_{\rm NRMSE} < 0.7$ indicate correct spectral shape with wrong amplitude, consistent with a missing normalisation factor or unit conversion error.

4.2.2. Deep Research Workflow Metrics

The Deep Research workflow produces no automated reference CSV, so ESR, PAS, and NAS are not applicable. Each task is instead evaluated across two dimensions: a quantitative parameter recovery score and a qualitative physical consistency assessment.

Parameter Recovery Score (PRS). Measures agreement between the agent's reported posterior median and the literature reference for each key parameter. For parameter p with reference $\theta_p^* \pm \sigma_p^*$, the per-parameter score is:

$$r_p = \max\left(0,\ 1 - \frac{|\hat{\theta}_p - \theta_p^*|}{3\,\sigma_p^*}\right), \quad (8)$$

saturating to zero at 3σ deviation. PRS is the unweighted mean of r_p over all task parameters.

Qualitative Assessment. Beyond parameter recovery, each task is assessed along two dimensions reported as free-text observations in the results: (i) *physical plausibility* — whether reported parameter values and posterior shapes are consistent with established physical expectations for the system; (ii) *failure transparency* — whether the agent identifies and reports known degeneracies, structural biases, or convergence issues, rather than delivering results silently — reported alongside PRS in the results.

4.3. Task-Level Robustness Evaluation

To evaluate reliability beyond aggregate performance, we design the research-driven task suite as a structured stress

test of the Deep Research workflow. Rather than ablating architectural components, we ablate the *task conditions* themselves—varying the degree of prior specification, likelihood completeness, and model complexity—while keeping the workflow fixed. This approach is motivated by prior observations that agentic failures in scientific settings emerge not from random errors but from systematic weaknesses under specific inference regimes: two structurally distinct under-constrained posteriors and compositional reasoning chains (Yehudai et al., 2025; Chowa et al., 2026).

We evaluate the One-Shot workflow through a controlled architectural ablation: access to domain-specific CAMB documentation is removed in the *CMBAgent (no context)* variant, isolating the contribution of structured retrieval augmentation on tool invocation accuracy and parameter configuration relative to both the full CMBAgent system and the direct Base LLM baseline.

The four research-driven tasks are assigned to these conditions as follows:

- **Under-Constrained Inference Stress Tests — Tasks 1, 2**

  Task 1: Fitting the Union2.1 Type Ia supernova distance–redshift data with a flat ΛCDM cosmology using MCMC to jointly estimate H_0 and Ω_Λ. The SN1a likelihood is degenerate in H_0 and the supernova absolute magnitude M_B; without an independent distance anchor, H_0 cannot be constrained by this dataset alone.

  Task 2: Modeling the rotation curve of NGC 3198 by combining stellar, gaseous, and bulge contributions with an NFW dark matter halo profile to infer halo virial mass M_{200} and concentration c, including uncertainty quantification via MCMC.

  Category Rationale: Both tasks probe silent overconfidence under known parameter degeneracies: the H_0–M_B degeneracy in T1 and the mass–concentration degeneracy in T2. A reliable agent should flag these as prior-dominated rather than report confident but physically uninformative constraints.

  Task label (T1): *Under-constrained* — flat prior on H_0, H_0–M_B degeneracy unbroken by design.

  Task label (T2): *Under-constrained* — flat priors on $(\log_{10} M_{200}, c)$, mass–concentration degeneracy unbroken by design.

- **Compositional Bayesian Workflow Evaluation — Tasks 3, 4**

  Task 3: Performing a hierarchical MCMC analysis of the exoplanet mass–radius relation for planets with $M > 2\,M_E$, incorporating intrinsic scatter, measurement uncertainties, and temperature-dependent radius inflation, with comparative analysis of Neptunian and Jovian regimes.

  Task 4: Conducting Bayesian inference for SLACS strong-lensing systems using a Singular Isothermal Sphere (SIS) model to recover posterior velocity dispersions and compare lensing-inferred values to observed stellar velocity dispersions.

  Category Rationale: Both tasks test silent inconsistency rather than degeneracy detection: whether the agent sustains physical consistency and stable parameterisation across a multi-stage pipeline with no artificial degradation. Failures here manifest as inter-trial drift, prior boundary pathologies, or population-level bias — none of which trigger obvious error signals.

  Task labels (T3, T4): *Compositional* — full hierarchical likelihood, no prior degradation.

5. Experiments

5.1. Systems

We evaluate three systems under the One-Shot workflow. The two CMBAgent variants are evaluated under their default configuration; full details of the agent-to-model assignments are provided in the CMBAgent repository (Xu et al., 2025). The Base LLM baseline uses a direct single-turn call to `GPT-4o-mini` with no surrounding architecture, providing a clean lower bound on performance attributable to the model alone.

- **Base LLM.** A direct single-turn call to `GPT-4o-mini` with no agentic framework, no tool integration, and no CAMB documentation. The model receives the task prompt and must produce executable Python code in one pass.

- **CMBAgent (no context).** The CMBAgent One-Shot workflow under default configuration, with the `engineer` agent and up to 50 reasoning rounds, but without access to CAMB documentation.

- **CMBAgent (CAMB context).** The CMBAgent One-Shot workflow under default configuration, with the `camb_context` agent providing retrieval access to the CAMB API documentation during execution. This is the full intended deployment mode of the system.

The Deep Research workflow is additionally applied to the four Astrophysical Research-Driven Tasks (T1–T4); its configuration and evaluation protocol are described in Section 5.2.

5.2. Protocol

Trials. The One-Shot workflow is evaluated over $N = 10$ independent trials per task–system combination, yielding $14 \times 10 \times 3 = 420$ total trials. All Base LLM calls use sampling temperature $T = 0.2$. The Deep Research workflow is evaluated with $N = 5$ independent trials per task; evaluation of this workflow is therefore qualitative, based on parameter recovery and physical consistency of the reported outputs rather than statistical aggregation across trials.

Ground truth. For the One-Shot workflow, reference outputs are generated once per task by executing the reference implementation from the CMBAgent benchmark repository (Contributors, 2024) and saving the resulting (x, y) numerical array to a standardised CSV file. For the Deep Research workflow, no automated reference CSV exists; ground truth is instead taken from the published literature values listed in Table 1, covering all four research-driven tasks (T1–T4).

Table 1. Literature ground-truth values for Deep Research evaluation.

Task	Param.	Reference	Source
T1	Ω_Λ	0.72 ± 0.02	(Suzuki et al., 2012)
T2	$\log_{10}\left(\frac{M_{200}}{M_\odot}\right)$	11.97 ± 0.15	(Karukes et al., 2015)
	c	6.8 ± 2.0	(Karukes et al., 2015)
T3	α_N	0.67 ± 0.05	(Müller et al., 2024)
	α_J	-0.06 ± 0.07	(Müller et al., 2024)
	$M_{\rm br}$ $[M_\oplus]$	127 ± 17	(Müller et al., 2024)
T4	f	1.019 ± 0.008	(Bolton et al., 2008)

Output standardisation. To ensure comparable output formats across systems, the target output file path is appended to every prompt at runtime. This eliminates ambiguity in output location without modifying the scientific content of the task prompt.

Deep Research configuration. Each Deep Research trial is allocated up to 50 planning rounds, 100 control rounds, and a maximum of 5 execution retries per sub-step (`max_n_attempts=5`). The retry budget allows the workflow to self-recover from environment compatibility errors such as deprecated API calls, which were the dominant failure mode observed across tasks.

Evaluation. For the One-Shot workflow, each trial output is evaluated against the reference CSV using the ESR, PAS, NAS, and failure mode metrics defined in Section 4.2. Per-task scores are averaged across 10 trials; per-system scores are averaged across tasks and trials. For the Deep Research workflow, evaluation is conducted by comparing the agent's reported posterior medians against the literature values in Table 1, with success additionally confirmed by the presence of output files beyond the execution log.

6. Results

6.1. One-Shot Workflow

Figure 1 summarises mean ESR, PAS, NAS, and Final Score across all 14 tasks and 10 trials per system.

CMBAgent with CAMB context. The full system achieves near-ceiling performance: ESR $= 0.96$, PAS $= 0.95$, NAS $= 0.86$, Final Score $= 0.85$. The small gap between PAS and NAS reflects residual numerical errors on the hardest tasks even when parameters are correctly configured.

CMBAgent without context. Removing documentation access produces sharp degradation across all metrics (ESR $= 0.62$, PAS $= 0.54$, NAS $= 0.18$, Final $= 0.15$). The most severe drop is in NAS, consistent with Figure 3 where Mode C (wrong computation) accounts for $\approx 47\%$ of trials — the agent invokes plausible but incorrect API calls, producing curves with the right shape but wrong amplitude or spectral content.

Base LLM. The direct `GPT-4o-mini` baseline fails on virtually all trials (ESR $= 0.09$, Final ≈ 0), with $\approx 91\%$ classified as Mode A (code failure). Raw model capability without agentic scaffolding is insufficient for tool-grounded scientific computation.

Task-level analysis. Figure 2 breaks performance down across all 14 tasks. CMBAgent with CAMB context scores near 1.0 on most tasks but fails silently on two: Task 10 (score 0.09), where most trials omit `raw_cl=True` producing a $\sim 10^5$ amplitude error and all trials omit the tensor B-mode contribution; and Task 14 (score 0.02), where the agent correctly computes per-ℓ delensing efficiency but collapses it to a scalar mean, yielding $CCC \approx 0$. Both failures stem from output construction errors rather than incorrect API usage — the documentation does not prescribe unit conventions or post-processing for these quantities. CMBAgent without context shows an isolated peak at Task 8 (0.63), whose calling convention is common enough to be known without documentation, but scores below 0.30 elsewhere. The Base LLM scores near zero throughout.

6.2. Deep Research Workflow

The Deep Research workflow was evaluated on four research-grade tasks (T1–T4) with up to five independent trials per task. Each task is reported with its PRS score and a qualitative assessment across the two dimensions defined in Section 4.2: physical plausibility and failure transparency. Table 2 summarises results.

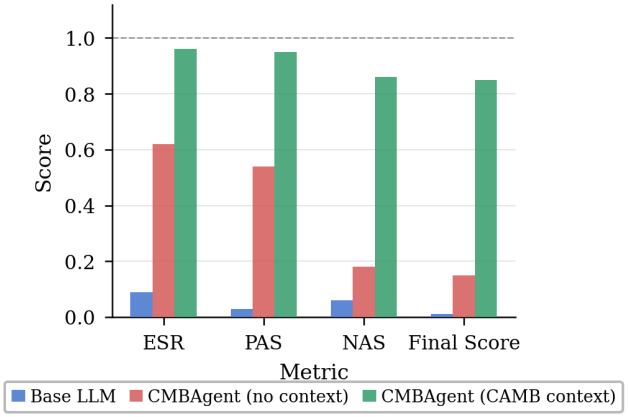

Figure 1. Mean ESR, PAS, NAS, and Final Score per system across 14 tasks and 10 trials.

T1 — SN1a Under-Constrained. All four completed trials recovered Ω_Λ within 0.15σ of the Union2.1 reference 0.720 ± 0.020 (Suzuki et al., 2012), with mean posterior 0.722 ± 0.001 confirming high reproducibility (PRS $= 0.97$). However, all trials deviated from the task specification by treating H_0 as a free parameter alongside Ω_Λ. This mirrors a realistic failure scenario: the prompt was intentionally constructed as a naive but plausible scientific request, and a reliable agent should have identified the H_0–M_B degeneracy and flagged that H_0 cannot be constrained from this dataset alone. Instead, the recovered H_0 posterior simply reflects the flat prior $H_0 \in [50, 90]$ rather than any data constraint, and was reported as a genuine measurement.

Physical plausibility: Ω_Λ posteriors are well-constrained and physically reasonable; the H_0 posterior is prior-dominated and physically uninformative.

Failure transparency: two silent failures on completed trials — the unauthorised addition of H_0 as a free parameter yielding a prior-dominated posterior, and a runtime constraint violation — neither flagged by the agent.

T2 — NGC 3198 Under-Constrained. All five trials failed to recover parameters consistent with the reference value (PRS $= 0.05$). Mass estimates spanned $\log_{10}(M_{200}/M_\odot) \in [10.0, 12.9]$ — nearly three orders of magnitude — with no trial within 5σ of the reference 11.97 ± 0.15 (Karukes et al., 2015). Three trials produced unphysical NFW concentrations: two high-mass trials ($\log_{10} M_{200} \approx 12.8$–$12.9$) returned $c < 2$, while one trial hit the upper prior boundary at $c \approx 30$.

Physical plausibility: three of five trials returned concentrations outside $c \in [2, 30]$; the unphysical $c < 2$ cases coincided with high virial mass, tracing a degenerate high-mass, low-concentration ridge.

Failure transparency: one trial noted elongated contours but no trial flagged unphysical posteriors or added caveats to its parameter estimates.

T3 — Exoplanet Mass–Radius Compositional. The Neptunian slope $\alpha_N \approx 0.60$ is consistently recovered across all four trials ($\sim 1.4\sigma$ from reference 0.67 ± 0.05, (Müller et al., 2024)). The Jovian slope α_J and break mass $M_{\rm br}$ are inconsistent across trials: Trial 0 recovers $M_{\rm br} \approx 127\, M_\oplus$ (within 1σ of reference) with $\alpha_J = 0.002$ ($\sim 0.9\sigma$); Trial 1 returns $\alpha_J = 0.368$ (6.1σ from reference -0.06 ± 0.07) and $M_{\rm br} \approx 203\, M_\oplus$ (4.5σ above reference); Trial 3 adopts a structurally different parameterisation with $\log M_b = 4.94$ on an inconsistent unit scale, precluding direct comparison; Trial 4 recovers $M_{\rm br} \approx 118\, M_\oplus$ (0.6σ below reference) with $\alpha_J \approx 0.000$ (0.9σ), but with $\sigma_{\rm int}$ (the intrinsic scatter in the mass–radius relation beyond measurement noise) pinned at the prior boundary (0.5), indicating inadequate prior specification. Mean PRS over the three key parameters gives PRS $= 0.73$, driven by robust α_N recovery but large variance in α_J and $M_{\rm br}$.

Physical plausibility: α_N is consistently physical; α_J is wrong in sign or magnitude in three of five trials; $M_{\rm br}$ is implausible in at least two trials (Trial 1 at 4.5σ above reference; Trial 3 on an inconsistent unit scale); $\sigma_{\rm int}$ hits the prior boundary in Trial 4.

Failure transparency: no trial flagged the inconsistent model parameterizations, implausible Jovian slopes, discrepant break masses, or the prior-boundary pathology in $\sigma_{\rm int}$; confident estimates were reported throughout without caveats.

T4 — SLACS Strong Lensing Compositional. Only one of five trials produced usable output; the remaining four failed to complete. The single completed trial yielded physically coherent results. However, lensing-inferred dispersions are systematically $\approx 5\%$ above observed values, yielding $f \approx 1.05$ against the reference $f = 1.019 \pm 0.008$ (Bolton et al., 2008), a $\approx 4\sigma$ bias (PRS $= 0.00$).

Physical plausibility: velocity dispersions and Einstein radius predictions are within the physically expected range and the SIS model is correctly implemented.

Failure transparency: the systematic offset in f was not flagged by the agent; the four failed trials produced no output or error diagnosis.

7. Conclusion

We presented a structured empirical evaluation of CMBAgent across two workflow paradigms and eighteen tasks spanning tool-grounded computation and research-grade Bayesian inference. Rather than focusing on aggregate success, our goal was to characterise *how* and *why* agentic systems fail, and whether those failures are detectable or

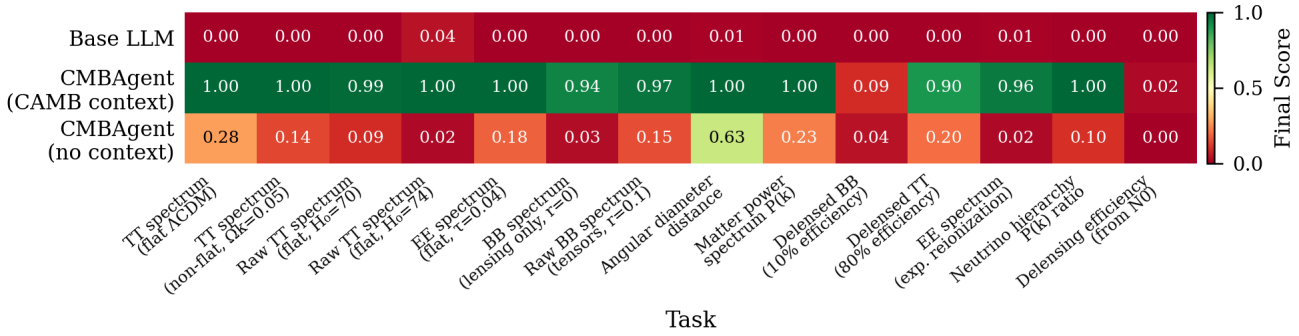

Figure 2. Final Score per system and task, averaged over 10 trials. Green indicates high score, red indicates low score.

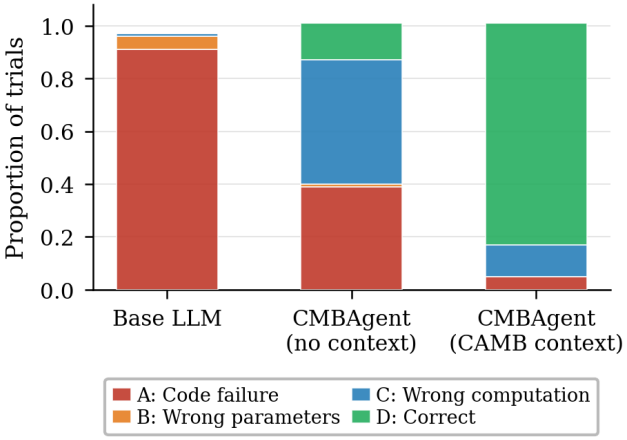

Figure 3. Failure mode breakdown per system as proportion of trials.

Table 2. Deep Research results. PRS is computed from posterior medians vs. literature for completed trials only; qualitative dimensions (PP = physical plausibility, FT = failure transparency) are rated ✓ (pass), ∼ (partial), or × (fail).

Task	Category	Trials	PRS	PP	FT
T1 SN1a	Under-constrained	4/5	0.97	∼	×
T2 NGC 3198	Under-constrained	5/5	0.05	×	×
T3 Exoplanets	Compositional	5/5	0.73	∼	×
T4 SLACS	Compositional	1/5	0.00	✓	×

silent.

Across both workflows, a consistent pattern emerges. In the One-Shot setting, domain-context retrieval is the dominant performance driver: with an identical `GPT-4o-mini` backbone, final score improves from ≈ 0 (Base LLM) to 0.15 (no context) to 0.85 (full context), a $\sim 6\times$ gain. Without context, the primary failure mode is not execution failure but silent wrong computation—syntactically valid code that produces plausible but incorrect results.

In the Deep Research setting, this failure pattern persists at the inference level. Under-constrained tasks fail systematically: degeneracies go undetected, yielding physically impossible posteriors, yet results are reported as valid. Compositional workflows execute successfully but exhibit inter-trial inconsistency, systematic bias, and unstable parameterisations without caveats. Across all tasks, outputs are consistently plausible, but failures remain unreported. Failure transparency is the weakest dimension: the agent never proactively flags known pathologies in its own outputs.

These findings highlight a critical risk for scientific deployment. Agentic systems do not primarily fail by crashing—they fail by producing confident, incorrect results or by silently breaking pipelines without diagnosis. Our evaluation framework, combining quantitative metrics for tool-grounded computation with structured qualitative assessment for scientific inference, provides a template for systematically identifying such failures. We argue that rigorous reliability evaluation is a prerequisite for deploying agentic AI in scientific workflows, where undetected errors can directly compromise scientific conclusions.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alansari, A. and Luqman, H. Large language models hallucination: A comprehensive survey. *arXiv preprint arXiv:2510.06265*, 2025.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Bolton, A. S., Burles, S., Koopmans, L. V., Treu, T., Gavazzi, R., Moustakas, L. A., Wayth, R., and Schlegel, D. J. The sloan lens acs survey. v. the full acs strong-lens sample. *The Astrophysical Journal*, 682(2):964–984, 2008.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D.,

Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chowa, S. S., Alvi, R., Rahman, S. S., Rahman, M. A., Raiaan, M. A. K., Islam, M. R., Hussain, M., and Azam, S. From language to action: a review of large language models as autonomous agents and tool users. *Artificial Intelligence Review*, 2026.

Contributors, C. Cmbagent benchmarks repository. https://github.com/cmbagent/Benchmarks, 2024. Accessed: 2026-03-02.

Ding, K., Yu, J., Huang, J., Yang, Y., Zhang, Q., and Chen, H. Scitoolagent: a knowledge-graph-driven scientific agent for multitool integration. *Nature Computational Science*, 5(10):962–972, 2025.

He, P., Dai, Z., He, B., Liu, H., Tang, X., Lu, H., Li, J., Ding, J., Mukherjee, S., Wang, S., et al. Traject-bench: A trajectory-aware benchmark for evaluating agentic tool use. *arXiv preprint arXiv:2510.04550*, 2025.

Hellert, T., Bertwistle, D., Leemann, S. C., Sulc, A., and Venturini, M. Agentic ai for multi-stage physics experiments at a large-scale user facility particle accelerator. *arXiv preprint arXiv:2509.17255*, 2025.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Karukes, E., Salucci, P., and Gentile, G. The dark matter distribution in the spiral ngc 3198 out to 0.22 rvir. *Astronomy & Astrophysics*, 578:A13, 2015.

Lin, L. I.-K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268, 1989.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

Lin, X. et al. LLM-based agents suffer from hallucinations: A survey of taxonomy, methods, and directions. *arXiv preprint arXiv:2509.18970*, 2025.

Moreno, E. A., Bright-Thonney, S., Novak, A., Garcia, D., and Harris, P. Ai agents can already autonomously perform experimental high energy physics. *arXiv preprint arXiv:2603.20179*, 2026.

Müller, S., Baron, J., Helled, R., Bouchy, F., and Parc, L. The mass-radius relation of exoplanets revisited. *Astronomy & Astrophysics*, 686:A296, 2024.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Qin, S., Zhu, Y., Mu, L., Zhang, S., and Zhang, X. Metatool: Unleash open-world function calling capabilities of general-purpose large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30653–30677, 2025.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., and Scialom, T. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

Sun, Q., Liu, Z., Ma, C., Ding, Z., Xu, F., Yin, Z., Zhao, H., Wu, Z., Cheng, K., Liu, Z., et al. Scienceboard: Evaluating multimodal autonomous agents in realistic scientific workflows. *arXiv preprint arXiv:2505.19897*, 2025.

Suzuki, N., Rubin, D., Lidman, C., Aldering, G., Amanullah, R., Barbary, K., Barrientos, L., Botyanszki, J., Brodwin, M., Connolly, N., et al. The hubble space telescope cluster supernova survey. v. improving the dark-energy constraints above z¿ 1 and building an early-type-hosted supernova sample. *The Astrophysical Journal*, 746(1):85, 2012.

Villaescusa-Navarro, F., Bolliet, B., Villanueva-Domingo, P., Bayer, A. E., Acquah, A., Amancharla, C., Barzilay-Siegal, A., Bermejo, P., Bilodeau, C., Ramírez, P. C., et al. The denario project: Deep knowledge ai agents for scientific discovery. *arXiv preprint arXiv:2510.26887*, 2025.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Xu, L., Sarkar, M., Lonappan, A. I., Zubeldia, Í., Villanueva-Domingo, P., Casas, S., Fidler, C., Amancharla, C., Tiwari, U., Bayer, A., et al. Open source planning & control system with language agents for autonomous scientific discovery. *arXiv preprint arXiv:2507.07257*, 2025.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., and Cao, Y. React: Synergizing reasoning and

acting in language models. In *The eleventh international conference on learning representations*, 2022.

Yehudai, A., Eden, L., Li, A., Uziel, G., Zhao, Y., Bar-Haim, R., Cohan, A., and Shmueli-Scheuer, M. Survey on evaluation of llm-based agents. *arXiv preprint arXiv:2503.16416*, 2025.

A. Deep Research Task Prompts

A.1. T1 — SN1a Under-Constrained

Read the file: `/home/sr/Desktop/code/cmbagent/cmbagent_systematics/deepresearch/task/SCPUnion2.1_mu_vs_z.txt`

Its description is: An ASCII table with tab-separated columns: Supernova Name, Redshift, Distance Modulus, and Distance Modulus Error. For Union2.1, there is an additional column for the probability that the supernova was hosted by a low-mass galaxy.

Fit this data within a flat ΛCDM model with two free parameters: H_0 and Ω_Λ. Write a simple but optimized MCMC to fit the SN1a data. Make a contour plot showing the 1D posteriors and quote the mean and 1σ on each parameter. Show the data alongside the best-fit model with 68% and 95% CL regions. Comment on the results.

Constraints: Running on a Dell Precision 5480 workstation with 32 GB RAM and 20 CPU threads (Intel Core i7-13800H). Use resources optimally so the MCMC converges within a few minutes. Have the engineer agent perform a preliminary MCMC timing step as a separate preliminary step.

A.2. T2 — NGC 3198 Under-Constrained

Read the galaxy rotation curve file: `/home/sr/Desktop/code/cmbagent/cmbagent_systematics/deepresearch/task/NGC3198_rotmod.txt`

Description: SPARC (Spitzer Photometry and Accurate Rotation Curves) provides high-quality rotation curves for nearby disk galaxies. For NGC 3198, the data includes: radius (kpc), observed rotation velocity (km/s), velocity uncertainty, gas contribution ($V_{\rm gas}$), disk contribution ($V_{\rm disk}$), and bulge contribution ($V_{\rm bul}$). The observed rotation curve traces the total gravitational potential.

Fit the rotation curve using an NFW (Navarro-Frenk-White) dark matter halo model combined with the baryonic contributions. The free parameters are: M_{200} (virial mass of the dark matter halo, in solar masses) and c (concentration parameter).

Write an optimized MCMC code to fit these parameters. Show the 1D posterior distributions and the 2D parameter contour plot. Quote the mean and 1σ values.

Plot rotation velocity versus radius showing: observed data with error bars, total best-fit model, individual components (disk, gas, bulge, dark matter halo), and 68% and 95% confidence bands.

Constraints: Run efficiently on a Dell Precision 5480 with 32 GB RAM and 20 CPU threads. Ensure the MCMC completes within a few minutes. Include a preliminary timing test.

A.3. T3 — Exoplanet Mass–Radius

Read the exoplanet data file: `/home/sr/Desktop/code/cmbagent/cmbagent_systematics/deepresearch/task/NASA_exoplanet_archive.csv`

Description: The NASA Exoplanet Archive provides confirmed exoplanet measurements. The file contains: planet name, mass (M_p in $M_\oplus$) with uncertainties, radius (R_p in $R_\oplus$) with uncertainties, equilibrium temperature ($T_{\rm eq}$ in K) with uncertainties, discovery method, stellar metallicity ([Fe/H]) with uncertainties, orbital period (days), semi-major axis (AU), and system distance (pc). Filter to planets with $M_p > 2\,M_\oplus$ to focus on the Neptunian and Jovian regime.

Fit a broken power-law mass–radius relation with two regimes separated by a break mass $M_{\rm br}$. Free parameters are: Neptunian slope α_N, Jovian slope α_J, break mass $M_{\rm br}$, normalisation, intrinsic scatter $\sigma_{\rm int}$, and a temperature-dependent radius inflation coefficient for hot planets.

Write an optimized MCMC sampler to obtain posterior distributions of all parameters. Implement: data filtering above $2\,M_\oplus$; hierarchical modelling of measurement uncertainties in both mass and radius; intrinsic scatter perpendicular to the mass–radius relation, and temperature-dependent corrections for radius inflation in hot planets.

Produce the following plots: mass–radius diagram with best-fit broken power-law curves; residuals versus mass showing the transition at $M_{\rm br}$; posterior distributions for α_N, α_J, and $M_{\rm br}$; a corner plot showing parameter correlations; and a separate comparison of hot planets ($T_{\rm eq} > 1000$ K) versus cold planets.

Constraints: Dell Precision 5480, 32 GB RAM, 20 CPU threads. Target runtime 5–8 minutes. Use analytic power-law models for efficient likelihood evaluation.

A.4. T4 — SLACS Strong Lensing Compositional

Read the galaxy lensing data file: `/home/sr/Desktop/code/cmbagent/cmbagent_systematics/deepresearch/task/Slac_data.csv`

Description: The SLACS (Sloan Lens ACS) survey provides high-quality data on strong gravitational lens systems. Columns include: z_d (lens redshift), z_s (source redshift), $R_{\rm eff}$ (effective radius, arcsec), $\theta_{\rm Ein}$ (Einstein radius, arcsec), $\sigma_{\rm obs}$ (stellar velocity dispersion, $\mathrm{km\,s^{-1}}$), and $\sigma_{\rm err}$ ($\mathrm{km\,s^{-1}}$).

Fit the lens mass profile using a Singular Isothermal Sphere (SIS) model with one free parameter per lens: the individual velocity dispersion $\sigma_{\rm SIS}$. Use z_d and z_s to compute angular diameter distances for each lens–source pair. Derive the model-predicted Einstein radius from $\sigma_{\rm SIS}$ within the SIS framework, and construct a likelihood comparing predicted $\theta_{\rm Ein}$ to observed values for each lens independently.

Write an optimized MCMC sampler to obtain the posterior distribution of $\sigma_{\rm SIS}$ for each individual lens.

Produce the following plots: 1D marginalized posteriors for representative lenses; a 2D comparison between lensing-inferred $\sigma_{\rm SIS}$ and observed $\sigma_{\rm obs}$ (one point per lens); model-predicted versus observed $\theta_{\rm Ein}$; and the mass profile $M(< R)$ for each lens overlaid with the Einstein radius scale.

Constraints: Dell Precision 5480, 32 GB RAM, 20 CPU threads. MCMC should converge within $\approx$5 minutes. Use vectorized likelihood evaluation across all lenses for efficiency.

B. Example Deep Research Outputs

B.1. T1 — SN1a Under-Constrained (Representative Trial)

The completed trial recovers $\Omega_\Lambda = 0.722 \pm 0.019$, consistent with the Union2.1 reference 0.720 ± 0.020 (Suzuki et al., 2012), but simultaneously reports $H_0 = 69.99 \pm 0.34\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ as a constrained result. The H_0 posterior is prior-dominated: the Union2.1 likelihood cannot break the H_0–M_B degeneracy without an external distance anchor. The agent did not flag this, presenting both parameters as equally reliable.

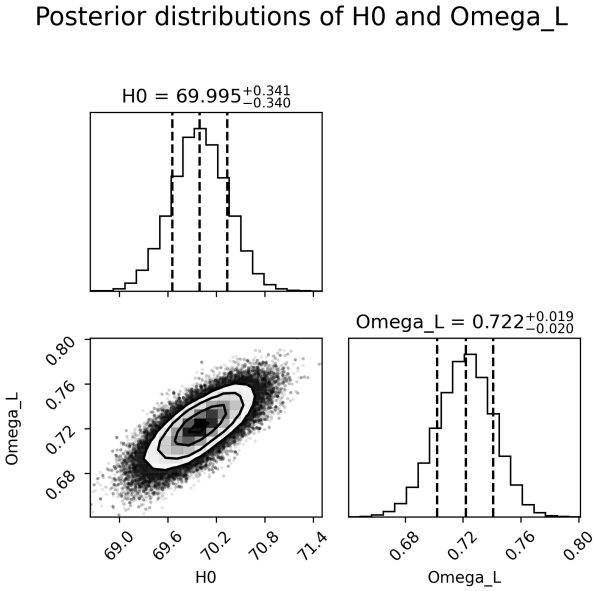

Figure 4. T1 posterior distributions of H_0 and Ω_Λ from the Union2.1 SN1a fit. While $\Omega_\Lambda = 0.722^{+0.019}_{-0.020}$ is well constrained and within 0.1σ of the reference, the H_0 posterior is prior-dominated — the Union2.1 likelihood cannot break the H_0–M_B degeneracy. The agent reported both parameters as equally reliable without flagging this pathology.

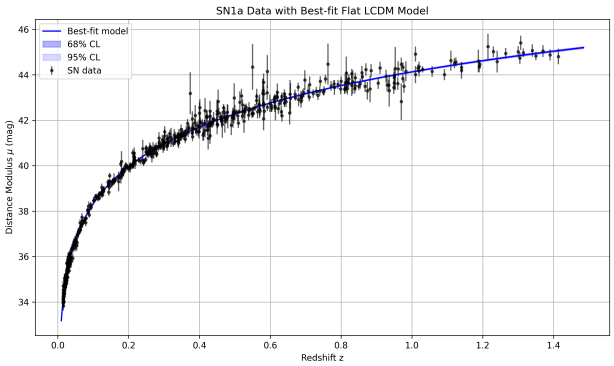

Figure 5. T1 best-fit flat ΛCDM model overlaid on Union2.1 SN1a data. The 68% and 95% CL bands are nearly indistinguishable from the best-fit curve, reflecting tight posterior constraints.

B.2. T2 — NGC 3198 Under-Constrained (Failure Case)

This trial illustrates the primary failure mode for T2: the posterior converges to $\log_{10}(M_{200}/M_\odot) = 12.90^{+0.06}_{-0.05}$, nearly 6σ above the reference 11.97 ± 0.15 (Karukes et al., 2015), with an unphysical concentration $c = 1.19^{+0.21}_{-0.14} \ll 2$. The agent reported these values without flagging the pathology.

B.3. T3 — Exoplanet Mass–Radius Compositional

This trial (Trial 3) illustrates the inconsistent parameterisation failure: $\log M_b = 4.939$ is reported on an inconsistent

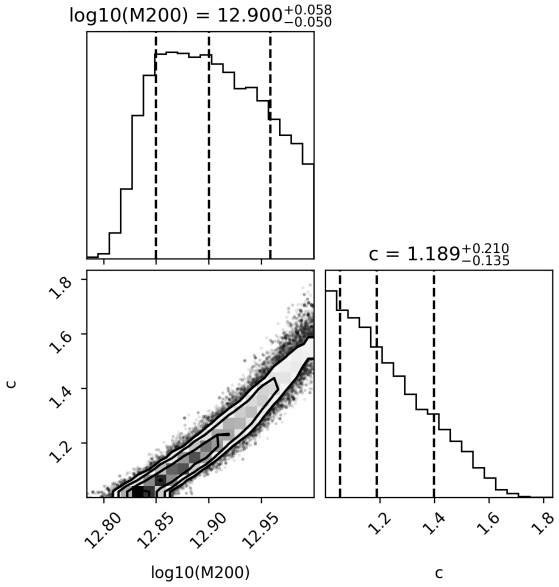

Figure 6. T2 posterior for $\log_{10} M_{200}$ and c, showing the degenerate high-mass, low-concentration ridge. The recovered $c \approx 1.2$ is physically impossible for an NFW halo; the reference value is $c = 6.8 \pm 2.0$ (Karukes et al., 2015). No trial flagged this pathology.

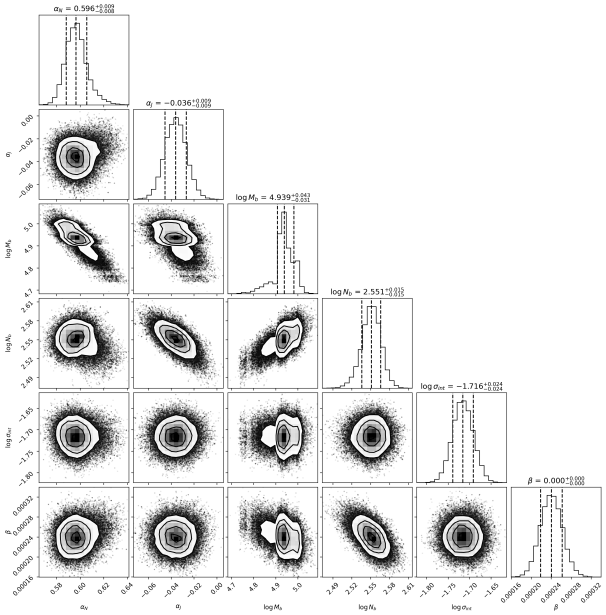

Figure 8. T3 corner plot from Trial 3. The break mass is reported as $\log M_b = 4.939^{+0.043}_{-0.031}$ on an inconsistent unit scale. The inflation coefficient $\beta \approx 0$ is effectively zero, indicating the temperature-dependent correction was not learned. No trial flagged these pathologies.

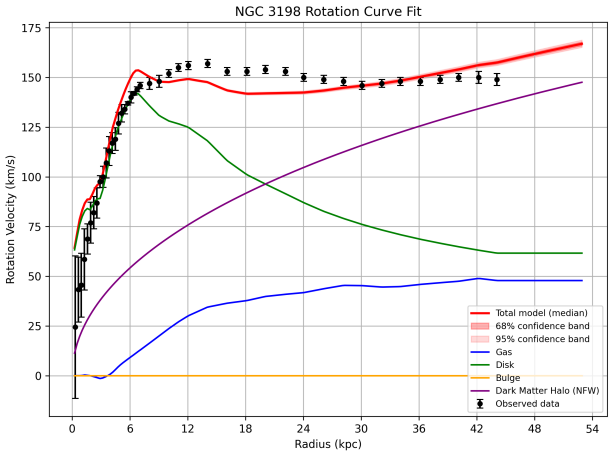

Figure 7. T2 rotation curve fit for NGC 3198. Despite a visually acceptable total model (red), the inferred NFW halo parameters are unphysical: the dark matter contribution (purple) is entirely dominant at all radii, inconsistent with the observed baryonic components.

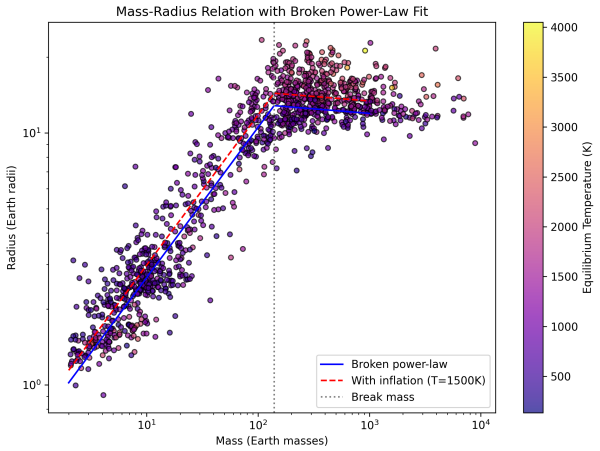

Figure 9. T3 mass–radius diagram with best-fit broken power-law. The fit appears visually reasonable, illustrating how plausible-looking outputs can mask underlying parameterization failures.

unit scale, precluding direct comparison to the reference $M_{\rm br} = 127 \pm 17\, M_\oplus$ (Müller et al., 2024). No caveat was generated by the agent.

B.4. T4 — SLACS Strong Lensing (Single Completed Trial)

Of five independent trials, only one produced usable output; the remaining four failed silently with no figures or error diagnosis. The single completed trial shows physically coherent velocity dispersions, but lensing-inferred values are systematically $\approx 5\%$ above observed values ($f \approx 1.05$ vs. reference 1.019 ± 0.008 (Bolton et al., 2008)), a $\approx 4\sigma$ bias the agent did not flag.

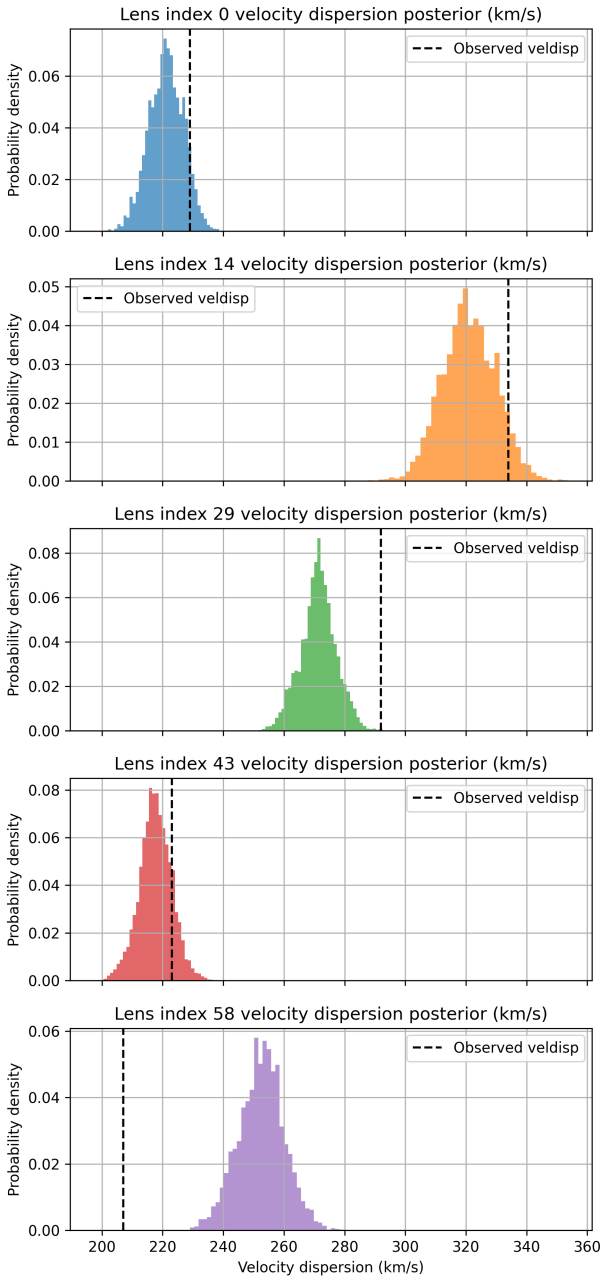

Figure 10. T4 representative 1D posteriors for five lenses. Several posteriors show tension with the observed stellar velocity dispersion (dashed line), consistent with the population-level $\approx 5\%$ systematic offset. No caveat was reported by the agent.

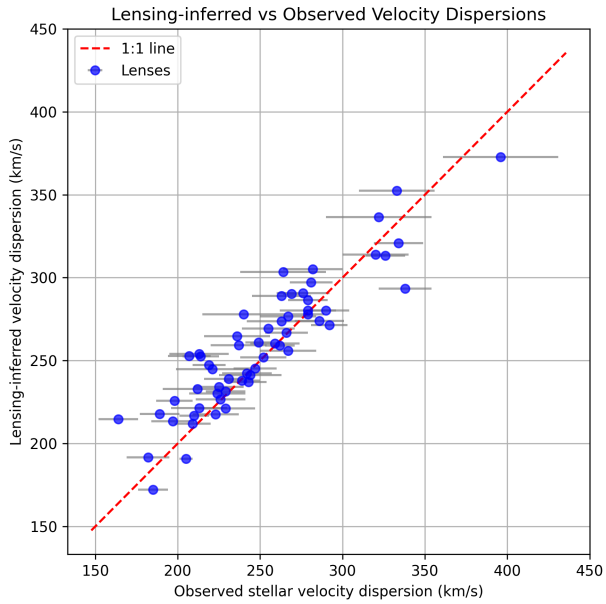

Figure 11. T4 lensing-inferred vs. observed stellar velocity dispersion $\sigma_{\rm SIS}$ vs. $\sigma_{\rm obs}$ (one point per lens). Points lie systematically above the 1:1 line (red dashed), yielding $f \approx 1.05$ against reference 1.019 ± 0.008 (Bolton et al., 2008).

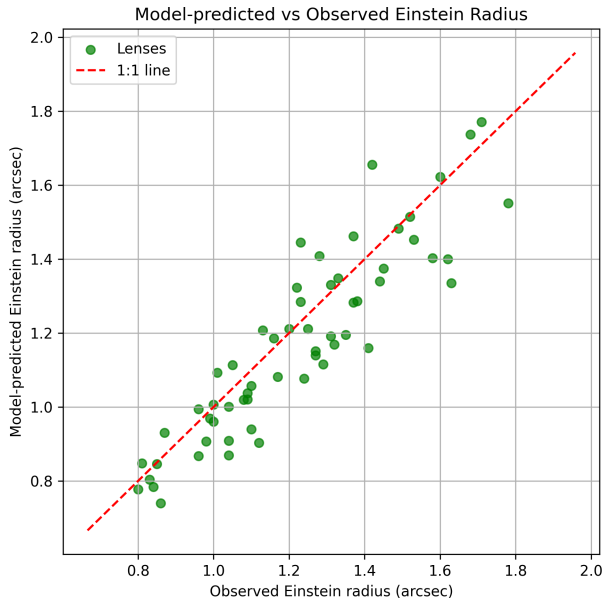

Figure 12. T4 model-predicted vs. observed Einstein radius $\theta_{\rm Ein}$. Agreement is closer to the 1:1 line here than for the velocity dispersions, consistent with the SIS model being correctly implemented but the overall calibration being offset.