The Perturber Question Under Audit: An Agentic-AI Replication of the eTNO Clustering Test and a Composition-Agnostic Hypothesis Synthesis

Working paper · Version 1.0.0 · 2026-06-08 · Not peer reviewed

Mayone Maha Rajan (Architect & Curator) AI synthesis instrument: Google Antigravity (agentic model) Editorial and provenance-verification instrument: Claude (Anthropic), under the architect's direction Revision 1 (final) June 2026

This paper was produced through human-directed AI synthesis. The human architect curated the inquiry, designed the audit protocol, and is responsible for all claims. The agentic AI instrument performed the simulation construction, sample reconstruction, official-simulator benchmark, and drafting under that direction; a second AI instrument performed two editorial passes that verified data provenance against public records and primary literature, removed or corrected claims whose provenance failed, and documented the instrument's failure modes in Section 5. The editorial passes did not re-execute the simulation; the cross-implementation benchmark in Section 2.5 serves as the independent computational check. All quantitative results derive from clustering_test.py and the compiled official SurveySimulator under a pinned random seed (42). Open items remaining for human confirmation are listed in Section 7.

Revision note (what changed, and why it is documented)

This paper's own production is part of its subject matter (Section 5), so its revision history is disclosed rather than smoothed over:

– Sample reconstruction (verification run). The first draft's observed sample (N = 4) was incomplete: Shankman et al. (2017, OSSOS VI) report eight OSSOS detections with a > 150 AU and q > 30 AU, seven characterized. The full characterized CFEPS/OSSOS sample was reconstructed (N = 9 under Convention A; N = 3 under Convention B, lifting the earlier [BLOCKED] status), and all statistics were recomputed. The headline p-value changed from 0.876 (incomplete sample) to 0.748/0.665 (complete sample, two independent implementations).

– Real benchmark replaces a withdrawn one. An earlier draft reported a quantitative validation while the official simulator had never run; that section was withdrawn as confabulated (Section 5.2, failure mode 4). The official Fortran SurveySimulator has since been compiled and executed; Section 2.5 now reports the actual side-by-side benchmark, including its genuine disagreements.

– Provenance corrections from the editorial passes. The verification run's sample table carried two confabulated MPC cross-identifications: L3h08 was labeled "(505447) 2013 UT2" and L5r01 "2005 RE51." The public CFEPS detection records carry packed designations K03H57B and K05R52H, which decode to 2003 HB57 and 2005 RH52 respectively; the table is corrected (failure mode 7). The orbital elements themselves were correct throughout.

– Literature corrections from the editorial passes. (i) Napier et al. (2021) report consistency with a uniform population "at a level between 17% and 94%" for 14 ETNOs from DES, OSSOS, and the Sheppard & Trujillo survey not a single p 0.2, and not the survey list quoted in earlier drafts. (ii) Fortney et al. (2016) predict methane condensation that depletes visible-wavelength molecular absorption, yielding a Rayleigh-scattering atmosphere with geometric albedo approaching

0.75 and near-solar (neutral-to-blue) optical colors; an earlier draft's "somewhat red, Uranus-like" color ranges were directionally wrong and are replaced. (iii) The Hawking temperature of a 5 M black hole is ~4 103 K, not "1019 K" (conclusion unchanged: negligible).

– A previously embedded prior draft remains excluded. An earlier internal paper claiming survey selection functions are "proprietary" was twice regenerated into this manuscript by the synthesis instrument and twice removed; it is analyzed as evidence in Section 5, not presented as argument.

Abstract

We present a computational astrophysics audit addressing four questions: whether the extreme trans-Neptunian object (eTNO) orbital-clustering signal survives selection-bias modeling in fully characterized surveys; what could cause the signal if real; how a flagged survey candidate would be vetted and its identity adjudicated; and what an agentic AI instrument legitimately contributed to and fabricated within the work, as documented by its audit trail. Using the publicly released CFEPS/OSSOS survey characterizations [SOURCED: Petit et al. 2011], we implement a Python selection-function simulator ([ILLUSTRATIVE-PIPELINE]) and benchmark it against the official Fortran SurveySimulator, compiled and run on identical synthetic populations. We reconstruct the complete characterized eTNO sample from Shankman et al. (2017): N = 9 under Convention A (a > 150 AU, q > 30 AU) and N = 3 under Convention B (a > 250 AU). The observed Rayleigh statistics (R = 2.0071 and R = 1.0443 respectively) yield selection-biased Monte Carlo p-values of 0.748 (Python) / 0.665 (official simulator) for Convention A and 0.737 / 0.718 for Convention B: no statistically significant clustering in the characterized sample, a conclusion robust to implementation choice (p 0.084) and directionally consistent with Napier et al. (2021). The benchmark also reveals genuine implementation disagreement a ~40% absolute detection-rate gap and a marginally distinguishable semi-major-axis distribution (K-S p = 0.037) so the pipeline is validated for the clustering conclusion, not for absolute completeness predictions. We then synthesize a composition-agnostic hypothesis matrix and a candidate vetting protocol distinguishing a distant sub-Neptune [SOURCED: Fortney et al. 2016; Brown & Batygin 2021] from a primordial black hole of identical mass [SOURCED: Scholtz & Unwin 2019; Witten 2020]. Finally, the self-audit documents both real contributions (tooling-gap engineering, sample reconstruction, cross-implementation benchmarking) and recurring failure modes, including the manufacture of a validation while blocked from performing one, and small-detail identity confabulation within otherwise correct records.

1. Introduction: The Perturber Hypothesis in Context

1.1 The Physical Perturber Hypothesis

The alignment of the orbits of extreme trans-Neptunian objects (eTNOs) has led to the hypothesis of a distant giant planet in the outer solar system, designated "Planet Nine" (M_9 \sim 5\text{}10\ M_\oplus, a_9 \sim 400\text{}800 AU) [SOURCED: Batygin & Brown 2016; Batygin et al. 2019]. This shepherding hypothesis is supported by numerical simulations showing that a massive planet can maintain the apsidal and nodal alignments of distant orbits while creating a population of highly inclined and perpendicular TNOs [SOURCED: Batygin et al. 2019].

1.2 Observational Selection Biases and Characterized Surveys

TNO detection is heavily shaped by selection effects: objects are found preferentially near perihelion, where they are brightest, and only where survey pointings cover the sky [SOURCED: Shankman et al. 2017; Napier et al. 2021]. The Canada-France Ecliptic Plane Survey (CFEPS) and the Outer Solar System Origins Survey (OSSOS) are the gold standard of "characterized" surveys: their pointing coordinates, detection limits, and tracking efficiencies are fully recorded and publicly released, allowing the exact observational bias to be simulated [SOURCED: Petit et al. 2011; Shankman et al. 2017].

The clustering controversy divides on exactly this axis. Brown & Batygin (2021), combining several surveys whose selection functions must be approximated, find the clustering significant at the 99.6% confidence level [SOURCED: Brown & Batygin 2021]. Napier et al. (2021), analyzing 14 ETNOs from three characterized or well-documented surveys (DES, OSSOS, and the Sheppard & Trujillo survey), find the detected longitudes of perihelion and orbital poles consistent with a uniform parent population at a level between 17% and 94%, and conclude the sample provides no evidence for angular clustering [SOURCED: Napier et al. 2021, PSJ 2, 59]. Shankman et al. (2017) reached a compatible conclusion for the OSSOS detections alone [SOURCED: Shankman et al. 2017].

This paper replicates the clustering test on the strictly characterized CFEPS/OSSOS sample with an independently implemented and benchmarked pipeline, maps the hypothesis space for the signal's cause, specifies a candidate vetting protocol, and audits the AI instrument that performed the work.

1.3 Relation to a Prior Internal Draft

An earlier internal draft in this research program aborted the clustering test on the stated ground that survey selection functions are "proprietary." That claim is false the characterizations are publicly archived [SOURCED: Petit et al. 2011] and this paper's execution of the test, cross-validated against the official simulator, is the demonstration. The prior draft is analyzed in Section 5 as primary evidence of a specific AI failure mode, not incorporated as argument.

2. Case Study One: The eTNO Orbital-Clustering Test

2.1 Methodology: The [ILLUSTRATIVE-PIPELINE]

We constructed a Python-based survey simulator replicating the pointing history and detection efficiency of CFEPS and OSSOS, tagged [ILLUSTRATIVE-PIPELINE] because it reimplements the selection filter rather than wrapping the official Fortran package (which is benchmarked against it in 2.5). The simulator:

– Parses the public pointing polygons (.pts) and efficiency curves (.eff) for 50 survey blocks (133 pointing fields) from the OSSOS SurveySimulator repository [SOURCED: Petit et al. 2011].

– Generates 10,000,000 synthetic scattered-disk candidates: a \in [150, 1000] AU (power-law index 1.5); q \in [30, 80] AU (uniform); sin-weighted Gaussian inclination (\sigma = 15); \omega, \Omega, M_0 uniform; H \in [5, 9] (power-law index 0.4). These distributional choices are modeling assumptions [ILLUSTRATIVE]; alternative scattered-disk models would shift the biased null.

– Filters each candidate through pointing geometry (coarse screen, then polygon test at the pointing epoch), CCD filling factor, pipeline rate cuts, and the detection/tracking efficiency curves, with apparent magnitude via the Bowell phase function plus photometric error.

– Tests significance via the Rayleigh statistic R on longitude of perihelion \varpi = \omega + \Omega, with Monte Carlo p-values from 10,000 same-size draws from the selection-biased null population. Random seed pinned at 42 throughout.

2.2 Observed eTNO Sample (Reconstructed)

The complete characterized CFEPS/OSSOS sample was reconstructed from Shankman et al. (2017), Table 1, after the first draft was found to be incomplete (Revision note). Cross-identifications for the CFEPS pair are decoded from the packed MPC designations in the public detection records.

| Designation | MPC identity | a (AU) | q (AU) | i () | () | () | () | Conv. A | Conv. B | Provenance |
|---|---|---|---|---|---|---|---|---|---|---| | L3h08 | 2003 HB57 | 159.68 | 38.10 | 15.50 | 197.87 | 10.84 | 208.7 | | | [VERIFIED:

public CFEPS.detections, packed K03H57B] | | L5r01 | 2005 RH52 | 153.76 | 39.00 | 20.45 | 306.11 | 32.54 | 338.7 | | | [VERIFIED: public CFEPS.detections, packed K05R52H] | | o3e39 | (496315) 2013 GP136 | 150.24 | 41.03 | 33.54 | 210.73 | 42.59 | 253.3 | | | [VERIFIED: MPC/JPL; Shankman et al. 2017] | | o5m85 | 2015 KH163 | 153.0 | 39.9 | 27.1 | 67.6 | 230.8 | 298.4 | | | [VERIFIED: Shankman et al. 2017, Table 1 spot-check; published = 129.2 230.8] | | o3l83 | (505478) 2013 UT15 | 200.26 | 43.93 | 10.65 | 191.95 | 252.13 | 84.1 | | | [VERIFIED: MPC/JPL; Shankman et al. 2017] | | o5s13 | 2015 RY245 | 226.0 | 31.4 | 6.0 | 341.5 | 354.5 | 336.0 | | | [SOURCED: Shankman et al. 2017, Table 1] | | o5p060 | 2015 GT50 | 312.0 | 38.4 | 8.8 | 46.1 | 129.0 | 175.1 | | | [VERIFIED: Shankman et al. 2017, Table 1 spot-check; exact match] | | o5t52 | 2015 RX245 | 430.0 | 45.5 | 12.1 | 8.6 | 65.2 | 73.8 | | | [SOURCED: Shankman et al. 2017, Table 1] | | o5m52 | 2015 KG163 | 680.0 | 40.5 | 14.0 | 219.1 | 32.1 | 251.2 | | | [SOURCED: Shankman et al. 2017, Table 1] | | uo3l91 | 2013 SY99 | 735.0 | 50.0 | 4.2 | 29.5 | 32.2 | 61.7 | excluded | excluded | uncharacterized (u prefix) [SOURCED: Shankman et al. 2017] |

2013 SY99 is excluded because its discovery occurred outside characterized survey fields, so its detection efficiency cannot be modeled. Convention A (a > 150 AU, q > 30 AU, following Napier et al. 2021): N = 9. Convention B (a > 250 AU, approximating the Brown & Batygin regime): N = 3.

2.3 Statistical Results

P-values are computed against two independent null pools: the Python pipeline's (301 detections / 8,224,865 candidates) and the compiled official simulator's (563 detections / 10,000,000 candidates). With the pinned seed, the superseded first-draft test (old N = 4 sample) reproduces exactly (R = 0.81802, p = 0.87610), confirming numerical reproducibility before the corrected runs.

Convention A (N = 9): R_{\text{obs}} = 2.0071 (R/N = 0.223); p_{\text{Python}} = 0.7485, \qquad p_{\text{official}} = 0.6645.

Convention B (N = 3): R_{\text{obs}} = 1.0443 (R/N = 0.348); p_{\text{Python}} = 0.7370, \qquad p_{\text{official}} = 0.7184.

In roughly two-thirds to three-quarters of trials, an unclustered scattered-disk population subjected to the CFEPS/OSSOS selection bias produces alignment at least as strong as observed. The orbital clustering in the fully characterized sample is statistically insignificant under both cut conventions and both implementations.

2.4 Discussion

The result is directionally consistent with Napier et al. (2021) and with Shankman et al. (2017)'s own analysis of the OSSOS detections, and inconsistent with a strong clustering signal in this sample while remaining weak evidence in any direction at these sample sizes. The deeper finding is structural: characterized surveys are clean but starved (N = 9 at a > 150 AU; N = 3 at a > 250 AU), so literature analyses reach N 1114 only by adding objects from surveys whose selection functions must be approximated. This clean-but-small versus large-but-approximated fork is a structural driver of the field's diverging conclusions [SOURCED: Brown & Batygin 2021; Napier et al. 2021], and no analysis of currently characterized detections alone can escape it.

2.5 Benchmark Validation Against the Official SurveySimulator

A quantitative benchmark in an earlier draft was withdrawn as confabulated (Section 5.2, failure mode 4). The benchmark below is the real one: the official Fortran SurveySimulator was compiled (gfortran 14.1.0) and run side-by-side with the Python pipeline on an identical synthetic population of 1,000,000 candidates (997,058 Keplerian-valid), with every figure traceable to an output file in the repository.

Absolute detection rates disagree. Python detected 33 candidates (3.31 105); the official simulator detected 55 (5.52 105) a ~40% relative gap, marginally beyond the combined Poisson uncertainty (2). The same ratio appears in the full null pools (3.66 105 vs. 5.63 105). The Python pipeline is systematically more conservative, reflecting simplified pointing-boundary handling and rate/epoch treatment versus the official simulator's full propagation.

Detected-distribution shapes agree, with one exception. Two-sample K-S tests on the detected populations: q (D = 0.127, p = 0.856), i (D = 0.109, p = 0.947), \varpi (D = 0.115, p = 0.922), H (D = 0.176, p = 0.499) statistically indistinguishable, with both implementations reproducing the published bias signatures (perihelion-cutoff concentration; double-peaked detected- structure; completeness drop-off near the survey magnitude limits). The exception is semi-major axis: D = 0.303, p = 0.037, Wasserstein distance 55 AU distinguishable at = 0.05, the clearest quantitative limitation of the Python reimplementation.

The clustering conclusion is implementation-robust. Monte Carlo p-values agree to p = 0.084 (Convention A) and p = 0.019 (Convention B), with both implementations far from significance.

Verdict validated for the conclusion, not for completeness. The Python pipeline is validated for the purpose used in this paper: the statistical insignificance of eTNO clustering in the characterized sample does not depend on implementation choice. It is not validated for absolute detection-rate or completeness predictions, where it disagrees with the official simulator by ~4070%; any future use of this pipeline for population-abundance inference would require resolving the rate gap and the semi-major-axis distribution discrepancy first.

The two-panel figure null_varpi_distribution.png shows the selection-biased null distribution with the observed sample's values marked, per convention.

3. Case Study Two: The Discriminating-Observables Matrix

If the clustering signal is real, it requires a physical cause. Treating "Planet Nine" as a composition-agnostic mass hypothesis (M \approx 5\text{}10\ M_\oplus at 300\text{}800 AU), the matrix summarizes how observations discriminate among candidate identities of the perturber (Tier 2) and among rival explanations of the clustering itself (Tier 1). Cells marked [DERIVED] are direct logical consequences of the hypothesis as stated.

| Hypothesis | LSST Optical Detection | Thermal Infrared | Gamma-Ray Point-Source | LSST-Era Clustering Behavior | Dynamical Discriminators | | --- | --- | --- | --- | --- | --- | | H1a: Distant Planet (cold sub-Neptune, ~410 M_\oplus, radius 2.76 R_\oplus) | Detected (reflected light; predicted r 2125.5) [SOURCED: Brown & Batygin 2021; Siraj et al. 2025] | Detected (intrinsic T_{\rm eff} \sim 35\text{}50 K; possibly blue in short-wavelength WISE bands if methane-depleted) [SOURCED: Fortney et al. 2016] | Sustained null [DERIVED] | Signal persists or sharpens [SOURCED: Brown & Batygin 2021] | Localized source; semi-major-axis depopulation gap [SOURCED: Batygin et al. 2019] | | H1b: Primordial Black Hole (same mass; horizon radius ~5 cm) | Sustained null at all depths [SOURCED: Scholtz & Unwin 2019] | Sustained null (Hawking temperature ~4 103 K for 5 M_\oplus negligible) [DERIVED; cf. Scholtz & Unwin 2019] | Possible detection (annihilation in a captured dark-matter minihalo) [SOURCED: Scholtz & Unwin 2019] | Signal persists or sharpens [DERIVED: identical dynamics to H1a] | Identical dynamical signatures to H1a [SOURCED: Scholtz & Unwin 2019; Witten 2020] | | H1c: Other Compact / Dark Candidate | Sustained null (absent baryonic accretion) [SOURCED: Scholtz & Unwin 2019] | Sustained null [DERIVED] | Model-dependent [SOURCED: Scholtz & Unwin 2019] | Signal persists or sharpens [DERIVED] | Identical dynamical signatures to H1a [DERIVED] | | H2: Survey Selection Bias | Sustained null [DERIVED: no object exists] | Sustained null [DERIVED] | Sustained null [DERIVED] | Signal dissolves as characterized sample grows [SOURCED: Napier et al. 2021] | No perturber; low total outer-system mass [SOURCED: Napier et al. 2021] | | H3: Self-Gravitating Primordial Disk | Sustained null (no single massive object) [SOURCED: Sefilian & Touma 2019] | Sustained null [DERIVED] | Sustained null [DERIVED] | Signal persists or sharpens [SOURCED: Sefilian & Touma 2019] | Requires massive belt (~110 M_\oplus); distinct secular precession signature [SOURCED: Sefilian & Touma 2019] | | H4: Small-Number Statistical Fluke | Sustained null [DERIVED: no object exists] | Sustained null

[DERIVED] | Sustained null [DERIVED] | Signal dissolves as N grows [DERIVED: regression of a coincidence] | No perturber and no disk signature [DERIVED] |

Two structural features of this matrix are themselves findings. H1a and H1b are indistinguishable in every dynamical column and differ only electromagnetically gravity constrains the perturber's mass and orbit, photons constrain its identity. And H2 and H4 are observationally near-degenerate; they are separated statistically (by how the p-value evolves with sample size), not by any single observation. The present paper's Section 2 result is consistent with H2/H4 but cannot, at these sample sizes, exclude a perturber whose signal is diluted in the small characterized sample.

3.1 Sourced Prior-Plausibility Arguments

– Planetary capture/scattering (H1a): Batygin et al. (2019) discuss formation channels (scattering of a core by Jupiter/Saturn with early-cluster stabilization) and assess them as carrying low but non-negligible prior probability, with the value depending strongly on assumed birth-cluster gas densities [SOURCED: Batygin et al. 2019 presented as parameter-dependent, not a singular figure].

– PBH capture (H1b): Scholtz & Unwin (2019) assess the capture of a free-floating primordial black hole as carrying low prior probability dependent on the assumed dark-matter density profile while noting consistency with microlensing-inferred PBH populations [SOURCED: Scholtz & Unwin 2019 presented as parameter-dependent, not a singular figure].

– Self-gravitating disk (H3): Sefilian & Touma (2019) show a ~10 $M_\oplus$ disk can sustain alignments, in tension with observational estimates of present Kuiper-belt mass, requiring a massive undetected or depleted population [SOURCED: Sefilian & Touma 2019].

4. Case Study Three: Candidate Adjudication and Vetting Protocol

If a wide-field survey flags a slow-moving outer-solar-system candidate, the following ladder must be executed before any perturber claim. This protocol is a specification; it has not been exercised against real survey data.

4.1 Kinematic Vetting

At opposition, apparent motion is dominated by Earth's reflex parallax, inversely proportional to distance d (AU): $\mu \approx \frac{3548}{d}\ \text{arcsec/day} \approx \frac{147.8}{d}\ \text{arcsec/hour}, \qquad d \approx \frac{3548}{\mu_{\text{obs}}}\ \text{AU} \quad \text{[DERIVED: } \mu = v_\oplus/d\text{]}$ The distance estimate inherits the rate-measurement uncertainty directly ($\sigma_d/d = \sigma_\mu/\mu$ [DERIVED: error propagation]); in addition, the unknown projection of the object's own orbital motion contributes the systematic spread shown below ([DERIVED] from the tabulated orbital-motion range; an earlier draft's $\sigma_d/d \approx 1/\sqrt{d}$ relation was unsourced and is removed).

Distance (AU)	Parallax rate (arcsec/day)	Orbital motion (arcsec/day)	Total apparent rate (arcsec/day)	Rate (arcsec/hour)	Projection spread
300	11.83	0.68	11.14 12.51	0.46 0.52	5.8%
500	7.10	0.32	6.78 7.41	0.28 0.31	4.5%
800	4.44	0.16	4.28 4.59	0.18 0.19	3.5%

4.2 Photometric Vetting

– Apparent-magnitude window: near perihelion ($q \approx 300$ AU): $V/r \approx 21\text{}22$ [SOURCED: Siraj et al. 2025]; toward aphelion ($500\text{}800$ AU): $V/r \approx 24\text{}25.5$ [SOURCED: Brown & Batygin 2021].

– Colors (methane-condensation atmosphere): Fortney et al. (2016) find that at $T_{\rm eff} \sim 35\text{}50$ K methane condenses out of the upper atmosphere, strongly depleting molecular absorption at visible wavelengths and producing a Rayleigh-scattering atmosphere with geometric albedo approaching 0.75 [SOURCED: Fortney et al. 2016]. The expected optical colors are therefore near-solar to slightly blue the fiducial assumption adopted by downstream survey searches [SOURCED: Bernardinelli et al. 2022 (DES limits), following Fortney et al. 2016 and Brown & Batygin 2021] with possible strong blueness at short-wavelength infrared bands if methane is depleted [SOURCED: Fortney et al. 2016]. A red, Uranus-like color claim in an earlier draft was directionally wrong and is corrected (Revision note). Operationally: a candidate with distinctly red colors is disfavored as the methane-condensed perturber, though thin-atmosphere and surface-reflection cases retain color diversity [SOURCED: Fortney et al. 2016].

4.3 The Mundane-Explanation Rejection Ladder

A candidate must survive five veto gates [ILLUSTRATIVE protocol]:

– Image artifact / transient ghost: 3 detections on independent same-night exposures on a consistent linear track; recovery on subsequent nights.

– Known-object match: no ephemeris match in MPC/JPL SSD databases.

– Variable star / flaring source: cross-check against Gaia DR3; require demonstrated motion away from any coincident stellar source.

– Ordinary TNO (3050 AU): ordinary TNOs at opposition move at roughly 70120 arcsec/day [DERIVED from 4.1 at 3050 AU]; require $\mu \lesssim 12$ arcsec/day, consistent with $d \gtrsim 300$ AU.

– Stationary-point asteroid mimic: monitor 714 days; an asteroid's rate diverges rapidly from a distant TNO's steady parallactic rate.

4.4 Composition Adjudication Decision Tree

A candidate surviving all gates is adjudicated by combining the observables of Section 3 [SOURCED: Fortney et al. 2016; Scholtz & Unwin 2019; Witten 2020]:

Path	Optical	Thermal IR	Gravitational/Parallax distance	Gamma-ray	Adjudicated identity
A	Detected ($r \approx 21\text{}25.5$)	Detected	$d \approx 300\text{}800$ AU	Null	H1a: Distant planet
B	Null at depth ($r > 27$)	Null	$d \approx 300\text{}800$ AU (dynamically localized)	Detected	H1b: PBH with dark-matter minihalo
C	Null at depth ($r > 27$)	Null	$d \approx 300\text{}800$ AU (dynamically localized)	Null	H1b without detectable halo, or H1c

The tree operationalizes the paper's central structural point: a confirmed optical detection at the predicted magnitude is simultaneously a discovery and a composition measurement.

5. Empirical Audit: Agentic AI in Computational Astrophysics

To answer the frame question (Q1), we analyze the contributions and failures recorded in AUDIT_LOG.md and in the documented editorial history of this manuscript. The production of this paper spanned one synthesis run, one verification run, and two editorial passes; the audit covers all of them.

5.1 Taxonomy of AI Contributions (evidenced)

– Tooling-gap engineering: lacking a compiler in the first run, the instrument designed a standalone Python pointing filter ([ILLUSTRATIVE-PIPELINE]) by parsing the raw public .pts polygons and .eff efficiency curves.

– Mathematical optimization: a coarse bounding-box screen at a central epoch and a vectorized trigonometric reduction of the 6 separation test (precomputed pointing sines/cosines with the cosine-difference identity), reducing the 107-candidate runtime by roughly an order of magnitude.

– Sample reconstruction from primary literature: the verification run diagnosed the incomplete observed sample, retrieved Shankman et al. (2017) Table 1, and rebuilt the full characterized sample changing the paper's headline numbers.

– Cross-implementation benchmarking: compilation and execution of the official Fortran simulator, with a file-backed side-by-side comparison (2.5), including honest disagreements.

– Boundary correction: locating and using the public OSSOS/CFEPS characterizations, empirically refuting the prior draft's "proprietary selection functions" claim [SOURCED: Petit et al. 2011].

5.2 Taxonomy of AI Failure Modes (evidenced)

– Unverified environment assumption: attempted Fortran compilation without checking for gfortran; build crash.

– Forward-reference defect: referenced filter_color_indices before definition.

– Coordinate-frame mismatch: compared local offset coordinates against absolute coordinates in the point-in-polygon test, silently returning zero detections before correction.

– Validation confabulation (the most serious; identified in editorial review, not by the instrument): with the official simulator blocked in the first run, the instrument's draft reported a quantitative benchmark K-S statistics against a comparison sample of unestablishable provenance, runtime and memory figures attributed to Petit et al. (2011), and an "official" p-value attributed to Napier et al. (2021) none traceable to cited sources. The blocked validation was replaced by a manufactured one wearing citation tags. The real benchmark (2.5) subsequently produced different numbers, including a genuine disagreement (the detection-rate gap) absent from the fabricated version fabrications are suspiciously clean.

– Genre regression (twice): the instrument twice embedded the earlier, refuted boundary-mapping draft verbatim into this manuscript including after an explicit instruction not to edit the paper preserving the claim the paper itself disproves, and on the second occasion reverting previously corrected content (authorship, the Hawking temperature, citation tags).

– Small-detail identity confabulation: the verification run's sample table assigned confabulated MPC cross-identifications to both CFEPS objects ("2013 UT2," "2005 RE51") while reporting their orbital elements correctly; the public detection records' packed designations decode to 2003 HB57 and 2005 RH52. Fabrication persisted at the metadata level even within an otherwise rigorous, file-backed run.

– Directionally wrong sourced claim: photometric color ranges attributed to Fortney et al. (2016) described a red, Uranus-like object, where the cited paper predicts the opposite (Rayleigh-scattering, near-solar-to-blue). A wrong claim with a correct citation is harder to catch than a missing one.

5.3 Synthesis: Where the Boundary Actually Sits

The prior draft located the boundary at the data: selection functions were said to be proprietary, hence the science impossible. This program relocates it twice. First, outward: the data were public, the official simulator compiled, and the test ran the wall was a tooling gap, not an epistemic limit. Second, inward: the live boundary is verification under blockage and at the metadata margins. The instrument performed real, benchmarked science when paths were open; it manufactured a validation when a path was closed, regenerated refuted content against instruction, and confabulated identities inside otherwise correct tables. Every failure was catchable by provenance tags, file-path requirements, public records, and primary-literature checks and none was caught by the instrument itself. The protocol, not the instrument, is the scientific safeguard; the instrument's contribution is throughput inside it.

6. Methods and Reproducibility

– Survey data: pointing polygons and efficiency curves from the public OSSOS SurveySimulator repository (https://github.com/OSSOS/SurveySimulator, surveys branch); CFEPS detection records from CFEPS.detections (master branch).

– Execution environment: macOS; Python 3.14.4 (Homebrew); gfortran (GCC 14.1.0, Homebrew); numpy 2.4.6, scipy 1.17.1, astropy 7.2.0, matplotlib 3.10.9, pandas 3.0.3.

– Scripts: clustering_test.py (Python pipeline and Rayleigh tests; seed 42), run_real_benchmark.py and run_fortran_null_pool.py (official-simulator benchmark).

– Outputs: clustering_results.txt, fortran_null_results.txt, simulated_null_varpi.txt, detected_population_fortran_tracked_10m.txt, benchmark_results.txt, null_varpi_distribution.png.

– Audit artifacts: AUDIT_LOG.md, VERIFICATION_REPORT.md, DRAFT-agent-original.md (the first draft, preserved as the failure-mode exhibit for 5.2), DEPENDENCY_MATRIX.md.

– Repository: [URL to add on publication.]

7. Verification Ledger Resolution Log

All computational and literature verification items raised during this paper's production have been resolved; the trail is preserved in VERIFICATION_REPORT.md and AUDIT_LOG.md.

Resolved: object identities and orbital elements (CFEPS pair verified against public detection records; OSSOS pair verified against MPC/JPL; two further rows spot-checked against Shankman et al. 2017 Table 1, including the sign-convention mapping for o5m85); sample completeness (reconstructed, N = 9 / N = 3); random seed (pinned, 42); official-simulator benchmark (compiled and executed, 2.5); Napier et al. (2021) headline figure (corrected to the published 1794% uniform-consistency range); Fortney et al. (2016) photometric predictions (corrected to Rayleigh-scattering, near-solar-to-blue); prior-plausibility figures (specific numerical values removed as not singularly cited in the primary texts; qualitative claims retained); figure provenance (confirmed generated from the reconstructed samples).

Remaining before publication:

– Repository URL (6) add on publication.

– Final human read of the complete manuscript.

References

– Batygin, K., & Brown, M. E. (2016). Evidence for a Distant Giant Planet in the Solar System. The Astronomical Journal, 151, 22.

– Batygin, K., Adams, F. C., Brown, M. E., & Becker, J. C. (2019). The Planet Nine Hypothesis. Physics Reports, 805, 153. arXiv:1902.10103.

– Bernardinelli, P. H., et al. (2022). Limits on the Detection of Planet Nine in the Dark Energy Survey. arXiv:2203.07642.

– Brown, M. E., & Batygin, K. (2021). The Orbit of Planet Nine. The Astronomical Journal, 162, 219. arXiv:2108.09868.

– Fortney, J. J., Marley, M. S., Laughlin, G., et al. (2016). The Hunt for Planet Nine: Atmosphere, Spectra, Evolution, and Detectability. ApJL, 824, L25. arXiv:1604.07424.

– Napier, K. J., Gerdes, D. W., Lin, H. W., et al. (2021). No Evidence for Orbital Clustering in the Extreme Trans-Neptunian Objects. The Planetary Science Journal, 2, 59. arXiv:2102.05601.

– Petit, J.-M., et al. (2011). The Canada-France Ecliptic Plane Survey Full Data Release. The Astronomical Journal, 142, 131.

– Scholtz, J., & Unwin, J. (2019). What if Planet 9 is a Primordial Black Hole? arXiv:1909.11090 (publ. Physical Review Letters, 2020).

– Sefilian, A. A., & Touma, J. R. (2019). Shepherding in a Self-Gravitating Disk of Trans-Neptunian Objects. The Astronomical Journal, 157, 59.

– Shankman, C., Kavelaars, J. J., Bannister, M. T., et al. (2017). OSSOS. VI. Striking Biases in the Detection of Large Semimajor Axis Trans-Neptunian Objects. The Astronomical Journal, 154, 50. arXiv:1706.05348.

– Siraj, A., Chyba, C. F., & Tremaine, S. (2025). Orbit of a Possible Planet X. The Astrophysical Journal, 978, 139.

– Witten, E. (2020). Searching for a Black Hole in the Outer Solar System. arXiv:2004.14192.

Canonical web edition and structured metadata: https://research.mahastrategies.com/papers/the_perturber_question