SCORING HOW, NOT WHETHER: A PROTOCOL FOR EVALUATING DISCOVERY IN SCIENTIFIC AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific research combines prior knowledge with evidence, and a discovery is usually made where the prior is absent or wrong. An agent can therefore contribute to discovery only if it derives claims from the evidence and lets that evidence override what it already knows or has been told. A correct answer alone cannot show which happened: reasoning from the evidence, recall from a prior, or acceptance of an answer it was handed. We therefore propose a protocol that evaluates the process separately from the outcome, combining controlled disclosure of the target identity, a planted false identity, and three trace measures: where each claim came from, whether it is backed by computed evidence, and whether a supplied identity was tested against the agent's own. The measures are read from the trace by judge models from outside the evaluated families, so that no model grades its own, and each is paired with a control condition or an input ablation. We validate the protocol on cancer subtype discovery with GPT-5.5, Sonnet 5, Opus 5 and Gemini 2.5 Pro. Four findings emerge. Withholding the cancer identity does not keep it hidden: GPT-5.5 and Sonnet 5 infer it from the gene symbols, Opus 5 names it before any gene is visible, most often from the sample count and clinical profile, and Gemini 2.5 Pro most often names none; the agent then confirms the cancer rather than considering alternatives, and the outcome score records none of this. The agents are better interpreters than analysts: all four propose well-supported mechanisms, yet once the identity is withheld three recover the published subtypes no better than a scripted k-means, and only Opus 5 beats it in every condition. A prescribed procedure improves the outcome score with no detectable reduction in adoption of a planted false identity. When a supplied identity conflicts with the agent's own evidence, Sonnet 5 reads its evidence as supporting the label in every episode it tests, GPT-5.5 records the contradiction and defers, Gemini rarely tests the label, and Opus 5 follows its evidence while accepting every true label; what separates them is which side prevails when a check fails. Human annotators agree with the judges on which side prevailed, less so on whether the label was tested. The results indicate that current agents are not yet reliable enough to be left unsupervised on a discovery problem: their answers can be recalled or accepted rather than derived, and a correct one does not show which. The protocol provides the means to examine this from the agent's log, and can be applied where no answer is available to check against.

1 INTRODUCTION

Autonomous agents built on large language models are increasingly applied to scientific tasks, and the benchmarks that evaluate them score the final answer. A correct answer, however, does not indicate how it was reached. An agent that derives the identity of a patient cohort from the molecular data in front of it and an agent that recalls the identity from the literature produce the same label and receive the same score, and the score does not register what either agent does when a premise supplied by the harness contradicts the evidence it has itself computed. Discovery requires both: deriving claims from the evidence, and letting that evidence overrule what is recalled or told. Recall is usually right where an answer exists, so a correctness score cannot tell the two apart. The distinc-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tion matters because such agents are being proposed for discovery pipelines and multi-agent systems (Gottweis et al., 2025), where no answer is available to check against; an agent whose answers are recalled or accepted rather than derived cannot be left unsupervised there, and nothing in its score reveals that it should not be.

To address these issues, we propose an evaluation protocol that scores the process of an agent separately from its outcome (Fig. 1). It consists of four instruments, each paired with a control condition that fixes its expected value in advance (§3): *provenance* and *support* scores read by a panel of judge families from outside the evaluated models; a *disclosure ladder* whose disclosed arm is a known-answer control in which recall is the expected reading; a *false-premise probe* that plants a plausible wrong identity through the harness's own channel and counts adoption only among episodes that received it; and a *premise-test judge*, blind to the truth and the verdict, that records whether the agent tested the supplied premise against its own evidence and which of the two prevailed.

Validation on a known-answer task. The process instruments do not require the answer (the support judge is handed a reference card for fact-checking in the validation run and is rerun without it, §A.7), and the one score that does require it, the outcome score, is the one unavailable in discovery. To show that the instruments measure what they claim, however, the study needs a task on which derivation and recall can be separated from one another: its answer must be known, so that the outcome can be scored, and published, so that recall is a genuine alternative to derivation. We therefore apply the protocol to blinded subtype discovery on seven TCGA cohorts (BRCA, LIHC, LUAD, LUSC, OV, PRAD and UCEC; 3,966 patients in all) (The Cancer Genome Atlas Research Network, 2013b; Hoadley et al., 2018), in which an agent must cluster patients, name the cancer and propose a mechanism from the data alone, across 760 episodes, four model families and seven cohorts.

Contributions.

- An evaluation protocol that scores the process of a discovery agent rather than only its answer (§3); applying it requires an elicited commitment, a logged harness and a plantable channel, and ground truth only to validate the instruments.
- Withholding the answer does not by itself produce derivation (§4.1): read from timestamps rather than a judge, 100 of 168 withheld identities are named only after the gene symbols are released, 30 are named before any symbol is visible, and 38 are never named; the outcome score cannot indicate which channel it scored.
- The agents are far better at interpreting their analysis than at performing it (§4.2): their mechanisms are rated grounded in 167 of 168 episodes, while their partitions recover the published subtypes no better than a scripted k-means (NMI 0.32 vs. 0.32, §4.2). This can be attributed to the fact that a mechanism can be recalled from the literature of a known cancer, whereas a partition must be derived from the data, the only route open in a discovery, where no such prior exists.
- Prescribing a procedure raises the outcome score by 9.5% and documented rigour by 20 points without changing false-premise adoption in any model (§4.3).
- When an agent's evidence conflicts with a supplied prior, the prior usually prevails (§4.4). Three families adopt a planted identity in $126/133$ exposed episodes; Opus 5 adopts in $15/51$. The difference is not whether the label is checked, which most models do, but whether the agent defers to the label or to its own evidence when the check fails.

2 Related Work

Scientific agents and their evaluation. End-to-end scientific agents have been described for chemistry, biology and machine learning (Boiko et al., 2023; Bran et al., 2024; Lu et al., 2024; Gottweis et al., 2025), and benchmarks evaluate discovery, data-science and computational-biology agents (Huang et al., 2024; Majumder et al., 2024; Jansen et al., 2024; Zhang et al., 2024; Chen et al., 2024; Mitchener et al., 2025); as in general agent benchmarks (Liu et al., 2023a; Xie et al., 2024b), they are scored on the correctness of their outputs. Several of these grade intermediate artifacts as well as answers, and process supervision (Lightman et al., 2023) and rubric-graded traces (Starace et al., 2025) score the steps of a solution, and biology benchmarks score answers or proposed experiments (Laurent et al., 2024; Roohani et al., 2024). Each of our instruments has such a neighbour: the

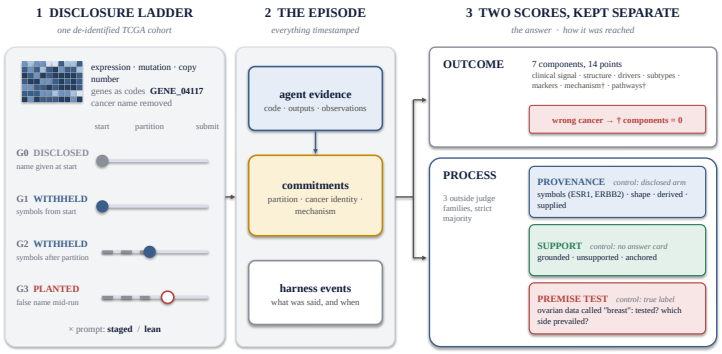


Figure 1: The evaluation protocol. *(1)* Disclosure ladder: one de-identified TCGA cohort (expression, mutation and copy number under opaque gene codes, cancer name removed) is run under four conditions: G0, disclosed, the true identity supplied at the start; G1, withheld, the identity absent but the gene symbols visible from the start; G2, withheld, the symbols released only after the partition is committed; and G3, planted, withheld with a plausible false identity asserted mid-run; each is crossed with a staged and a lean prompt. *(2)* The episode: the agent's evidence (code, outputs, observations), its commitments (partition, cancer identity, mechanism) and every harness event are written to the log with the call at which they occurred. *(3)* Two scores, kept separate: outcome, a seven-component, 14-point rubric computed from the submission, whose identity gate zeroes the two disease-interpretation components (†) on a wrong cancer; and process, read from the trace by three judge families from outside the tested vendors under a strict majority, comprising provenance (where the identity came from), support (whether a claim is backed by computed evidence) and the premise test (whether a supplied identity was tested, and which side prevailed), each paired with a control: the G0 arm, a rerun without the answer card, and a true label.

planted label is indirect prompt injection crossed with a context–knowledge conflict, the disclosure ladder is a contamination-resistant blinding design, and the process scores are trace rubrics. What is new is the composition: a commitment elicited from the agent, a logged channel through which a premise can be planted, a known-value control for every instrument, and an exposure-conditioned denominator, which together let derivation be told from recall where a trace rubric alone cannot.

Reasoning traces and model judges. A chain of thought can rationalise an answer without reporting the feature that caused it (Turpin et al., 2023; Lanham et al., 2023; Chen et al., 2025). We therefore treat traces as evidence to be checked, with a control per judge. Language-model judges prefer longer responses and their own family (Zheng et al., 2023; Liu et al., 2023b; Chiang & Lee, 2023; Dubois et al., 2024; Panickssery et al., 2024), so no judge in the panel belongs to an evaluated family.

Sycophancy, false premises and memorisation. Language models agree with a supplied belief even when it is wrong (Perez et al., 2023; Sharma et al., 2023; Wei et al., 2023), answer false-premise questions (Hu et al., 2023), are derailed by irrelevant context (Shi et al., 2023), resolve context–knowledge conflicts inconsistently (Xie et al., 2024a; Xu et al., 2024), and follow untrusted tool content (Zhan et al., 2024). The planted label here is a plausible assertion delivered through the trusted harness against contrary computed evidence, with whether the label was checked scored separately from which way the conflict was resolved.

3 The Protocol

The protocol has a production part (§3.1), what a task must provide and run, and an evaluation frame (§3.2), how each episode is scored and checked; §3.3 describes the implementation.

3.1 PRODUCTION PROTOCOL

R1. What a task must provide. To *apply* the protocol, a task needs a scored commitment that the agent is asked to make (R3), a harness that logs what it tells the agent (R4), and a *plantable channel* through which the harness can assert that commitment with authority. To *validate* the instruments, the task additionally needs ground truth, so that the disclosed condition can serve as a known-answer control, and a recallable prior, so that recall is a genuine alternative to derivation. Ground truth enters through that calibration, the outcome score and, in the validation run, a reference card handed to the support judge for fact-checking (removed in a rerun, §A.7); the provenance and premise-test judges are blind to it.

R2. Three conditions: the disclosure ladder and the false-premise probe. In the *disclosed* condition the commitment is supplied, true, at the start; derivation is unnecessary, which fixes the expected provenance reading (recall) and premise-test reading (acceptance) there; the support reading is checked instead by removing the judge's reference card (E4). In the *withheld* condition the commitment is not supplied, and the information from which it could be recognised is released only after the agent has committed to the structure it will later be asked to interpret. The *planted* condition is the withheld condition with a plausible false commitment asserted mid-analysis through the plantable channel.

R3. Elicit the commitment. The submission must ask explicitly for the scored quantity, as a field with a confidence; a model that answers exactly what it is asked will otherwise appear never to commit, and every provenance instrument downstream will be scoring its silence.

R4. Log what the harness said, and when. Every assertion the harness makes to the agent (the disclosed answer, the released vocabulary, the planted premise) is written to the episode log with the call index at which it was delivered, so that exposure can be reconstructed per episode; an episode in which the premise was never delivered cannot be counted as having resisted it.

3.2 EVALUATION FRAME

E1. Three scores per episode, kept separate: outcome, provenance and support. *Outcome* measures whether the answer is good, on a component rubric whose gate zeroes only the components that presuppose the scored commitment. *Provenance* records where the commitment came from, *explore* (earned from this episode's evidence) or *exploit* (taken from a prior or from the harness), with a finer *recall type* (derived, recalled then grounded, asserted, never committed). *Support* records whether the commitment is backed by evidence the agent computed here (grounded), unbacked (unsupported), or held against contrary computed evidence (anchored). A fourth descriptor, *analytic provenance* (whether the surrounding analysis was computed rather than recited), nests the provenance score and must not be substituted for it (§A.9).

E2. The premise-test judge. For every episode in which the harness supplied the commitment (true in the disclosed condition, false in the planted one), a judge reads the distilled trace plus one line stating what was supplied and when, and returns the *handling* (tested and rejected, tested and undecided, tested and accepted, accepted untested, ignored), the *direction* of the agent's own evidence (supports, contradicts, mixed, none computed), and the *final relation* of the submission to the supplied commitment. The judge does not see the truth, the condition or the outcome verdict. The disclosed condition is its control: testing a true premise is permitted, rejecting one is an error.

E3. Judges and consensus. Process scores are read by a panel of judge families from outside the evaluated vendors, one pass each. A label is assigned only on a strict majority; otherwise the episode is *no-consensus*, counted and reported, never folded into a label or dropped. The outcome gate is read by the same panel; its pooled rate varies by 1.2 points across families, with 112 episode-level disagreements resolved by majority and two exposed episodes unresolved (§A.8).

E4. Calibration checks and analysis rules. Every process instrument is read at the disclosed condition, where its value is fixed a priori; the primary disclosure contrast is recomputed under every subset of the panel, and where the magnitude varies with the subset the direction is the claim;

a judge handed the answer for fact-checking is rerun without it (§A.7); a contrast smaller than the seed range within a cell is not a finding; anchors of known quality pass through the same rubric; and a stratified sample of process labels is re-read by annotators blind to the panel (§A.11). In the analysis, rates on the planted condition are computed over episodes that received the premise, pooled contrasts are shown within each model before they are quoted, an empty scorer result is a failure rather than a zero, intervals are clustered on cohorts except where a figure states otherwise, per-model rows accompany every pooled number, and no-consensus counts are reported per field (§A.19).

3.3 Implementation on TCGA subtype discovery

The task is subtype discovery on a real, de-identified TCGA cohort, with data from the Genomic Data Commons, cBioPortal (copy number) and UCSC Xena (published subtypes), fetched by the package's scripts. Writing and running its own Python code, the agent commits to a partition of the patients (D1), an identity, that is, which cancer these patients have (D2), and a mechanistic hypothesis (D3). D2 is the scored commitment: it has a plantable channel, the codebook reveal, and, for validation, ground truth and a memorised prior for every public cohort. Episodes run offline with identity-bearing files unreadable; the agent sees expression, mutation and copy-number matrices under opaque `GENE_XXXXX` identifiers and five clinical columns, with sex and diagnosis removed. The conditions are G0 (disclosed), G2 (withheld; gene symbols released at the third recorded observation, which the staged prompt places after the partition is committed; the pre-reveal partition is not hashed, and re-clustering after the reveal is discouraged by the prompt rather than prevented) and G3 (planted; a false cohort label injected through the codebook channel at the third or fifth observation); G1 (gene names released immediately) and a staged-versus-lean prompt ablation are reported as case studies. Outcome is a seven-component rubric summing to 14 (Table 1), six components computed deterministically and one judged by a language model, gated on D2. Provenance, support and the premise test are read by `nemotron-3-super`, `laguna` and `Qwen3.6-27B`; no-consensus is $153/760$ on the D2 tag and $32/760$ on analytic provenance. The study comprises 760 episodes over four model families (GPT-5.5, Sonnet 5, Opus 5, Gemini 2.5 Pro; the second and third share a vendor) and seven cohorts (The Cancer Genome Atlas Research Network, 2013b; Hoadley et al., 2018), at three seeds in the honest arms and two source–label pairs at eight seeds in G3. Four Gemini 2.5 Pro lean-prompt episodes (three honest, one G3) submitted an empty partition and have no outcome score; they are excluded from outcome denominators as failures rather than scored as zero (E4) and kept in the process and identity analyses. One partition per episode, the one the scorer received, is used throughout (§A.3).

4 Experimental Results

Unless stated otherwise, intervals are cohort-clustered 95% bootstrap intervals, paired contrasts use the Wilcoxon signed-rank test, proportions carry Wilson intervals, and every rate on the planted condition is computed over exposed episodes (E4). We first establish how the outcome score itself behaves under these contrasts (Fig. 2). The composite score responds to the cohort ($+65\%$ between the hardest and the easiest) and to the scaffold ($+8\%$), and weakly, if at all, to the process contrasts the protocol measures: two rows, derived against recognised (-8%) and tested against untested (-5%), have intervals excluding zero. Mechanism grounding, the interpretive component, moves by less than 4% under any contrast, the cohort included. Subtype recovery, the partition component, does respond, and where it responds it favours recall: a derived identity scores 35% below a recognised one, a withheld identity 14% below a disclosed one, and an identity read from the shape of the dataset 46% above one read from the symbols. The composite therefore averages a component that cannot move with one that moves toward recall. On the un-gated components, resisting a false label is worth 6% with an interval through zero (30% on subtype recovery).

4.1 Effect of withholding the identity on provenance and outcome

The manipulation check of the ladder passes: withholding the identity raises the explore-tagged share of D2 decisions from 1.3% to 46.7% under the staged prompt and from 9.0% to 44.6% under lean, in every model $\times$ prompt lane and judge configuration, with magnitude from $+8.3$ to $+71.4$ pt across the 14 configurations, so the direction is the claim (§A.6). The identity decision responds

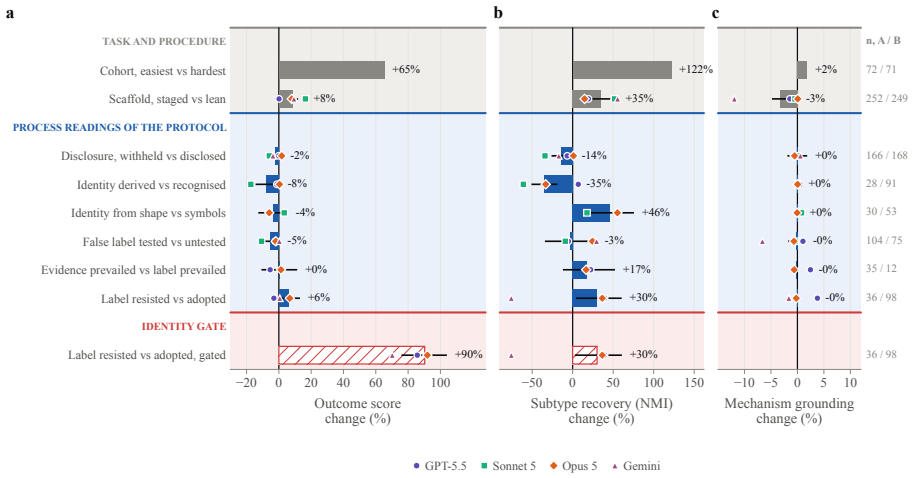


Figure 2: Sensitivity of the outcome score and of two of its components to task, procedure and process contrasts: *(a)* the composite outcome score; *(b)* subtype recovery, the NMI with the published subtypes (the partition component); *(c)* mechanism grounding (the interpretive component); rows shared. Each row is the change from group B to group A within model: the bar averages the within-model differences (weight $\min(n_A, n_B)$) over the pooled group-B mean, with 95% intervals from resampling episodes within model and group, not cohort-clustered; markers are individual models over the same denominator, omitted below three episodes per group; n at right is for the composite. Rows: the easiest cohort against the hardest and the staged against the lean prompt (honest arms); identity withheld (G2) against disclosed (G0); at G2, identities the panel reads as derived against recognised, and identities named before the gene symbols were released against after; on exposed G3 episodes, the false label tested against accepted untested, the evidence prevailing against the label prevailing where the two contradicted, and a final identity that resisted the label against one that adopted it, all un-gated. The bottom row repeats the last contrast on the gated composite, where the gate zeroes the two disease-interpretation components on a wrong identity; subtype recovery is not gated, and the gate sets the adopters' mechanism grounding to 0, so that cell is empty. The first three rows are experimental contrasts, the rest descriptive strata by judge label.

most (Fig. 9): the partition tag starts above 85% and rises by 2 to 18 points depending on the judge, and the mechanism tag is separated from zero in 3 of 14 judge configurations (§A.6, §A.19). The result concerns where the withheld identities came from, which the episode log answers without a judge by comparing when the agent first names a disease with when the gene symbols arrived (Fig. 3b, §A.12). Under G2, 100 of 168 episodes name the cancer only after the symbols are released; 38 never name one (31 of them Gemini); and 30 name it before any symbol is visible, 29 correctly and 24 of them Opus 5; 19 of the 30 cite the dataset's shape (sample count, deaths, age, stage) and the rest statistics on the opaque genes (§A.12). The judge panel calls 28 identities derived from computed markers and 94 recalled then grounded; 24 of the 28 are symbol-channel episodes, whose derivation ran on the released symbols, and 25 of the 30 shape-channel identities are read as recall (judge-dependent, §A.8). Recognition is not a failure of grounding: 91 of the 100 symbol-channel identities and all 30 shape-channel ones are grounded, the first cancer named in the trace is correct in 86 of the 95 episodes that name one before the final submission and in 29 of 30, and the submitted identity passes the gate in 94 of 100 and 30 of 30 (Fig. 3c). It is a failure of testing. The agent confirms the cancer it has recognised and does not ask what else the data could be (§4.4).

The outcome score does not distinguish these channels. Within each model the outcome is the same whether the identity was disclosed, recognised from the symbols or read from the shape (Fig. 2): the score awards no credit for a correct identity, its gate fires only when the identity is wrong (5 of 168 G2 episodes), and never-named episodes are not separated from the rest within any model (Fig. 3d). The balance between exploration and recall is therefore set by what the environment discloses rather than chosen by the agent.

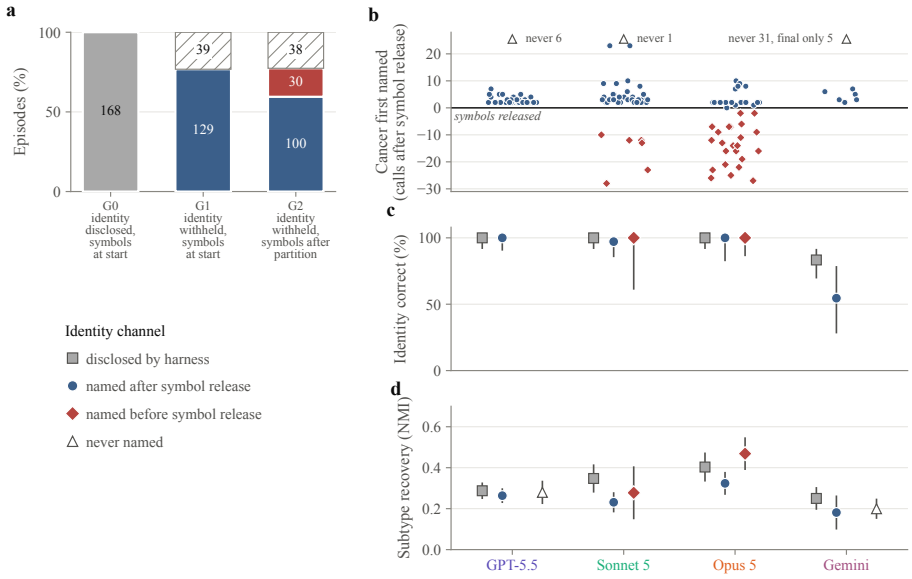


Figure 3: Source of the identity claim under each disclosure condition (conditions as labelled in *(a)*). *(a)* Per condition, where the named identity came from, read from timestamps: disclosed by the harness (grey), named after the gene symbols were released (blue), named before any symbol was visible (red), or never named (hatched). *(b)* At G2, per model, the call at which the agent first named a cancer, counted from the symbol release (0): red diamonds before, blue circles after; never-named and final-submission-only episodes are counted at the top. *(c)* Identity correctness by channel within model, Wilson intervals, with the model's G0 episodes as "disclosed". *(d)* NMI between the agent's partition and the published subtypes, by channel within model, mean $\pm$ 95% CI.

4.2 Quality of the analysis and of its interpretation

Even with the identity disclosed, the agents' partitions recover the published subtypes at a mean NMI of 0.32 (Table 16; Fig. 3d), which is the 0.32 that a scripted k-means on the top-variance genes reaches with no knowledge of the cohort (§A.2). On the interpretive components the same episodes are close to ceiling: mechanism grounding averages 0.93, pathway validity 0.68, and the support panel rates 167 of 168 mechanisms grounded. Knowing the cancer therefore improves the interpretation of the analysis and the choice of markers, not the analysis itself. NMI rises with the number of clusters, and the agents commit to between 2 and 9 groups (median 4; 126, 126, 155 and 71 episodes at $k = 2, 3, 4$ and 5), so we recomputed it with adjusted mutual information and against a k-means run at each episode's own k on the same seed (Table 3): the pooled tie holds (AMI 0.29 against 0.30), Opus 5 is above its k-matched baseline in every arm, Sonnet 5 at G0 only, and GPT-5.5 and Gemini below it in every arm. Neither interpretive component checks that a mechanism is the right one for the subtypes, only that it is evidence-backed (§A.2).

Mechanism grounding is at least 0.98 for GPT-5.5, Sonnet 5 and Opus 5 in every arm and 0.73–0.76 for Gemini (§A.2). The agents therefore write well-grounded interpretations of partitions a scripted baseline would match, and the rubric, which gives subtype recovery and mechanism grounding the same weight, rewards the one as readily as the other. Grounded and correct are different properties.

4.3 Effect of a prescribed procedure

The scaffold ablation pairs the staged and lean prompts by model $\times$ cohort $\times$ seed $\times$ arm. Over the 249 honest-arm pairs the staged prompt scores $+0.043$ higher, a 9.5% rise on the lean mean of 0.456 (CI $[+0.025, +0.062]$, positive in $7/7$ cohorts), redistributed toward the components that reward procedure while mechanism grounding falls (§A.17). Three of the four families carry it and GPT-5.5 is flat (Fig. 4a). Documented validation rigour rises with it, by 19.5 pt (CI $[+11.1, +26.8]$). The process readings are not a clean procedure contrast, because the staged prompt violates rule R3: it never asks for the cancer type, leaves 25 of 84 G2 episodes without an identity where the lean

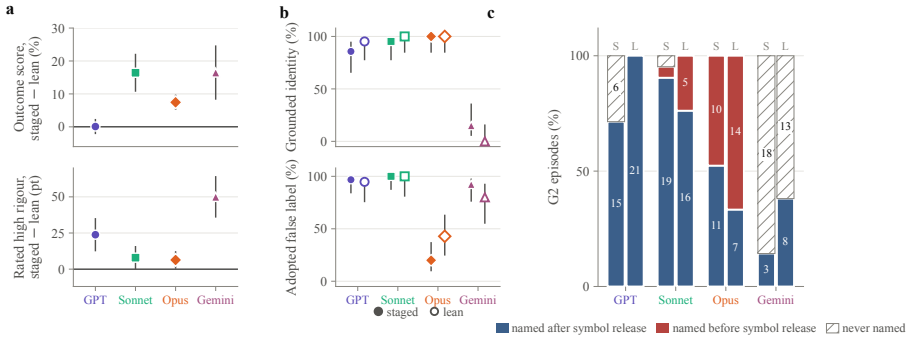


Figure 4: Effect of prescribing a procedure: the staged prompt against the lean prompt, paired within model. *(a)* Paired change under the staged prompt in the outcome score, as a percentage of the lean twin's mean, and in the share of episodes rated high in documented rigour, in points; mean ± 95% CI over pairs. *(b)* Under each prompt (filled staged, open lean), the share of G2 twins whose identity claim is grounded in the agent's own evidence and the share of exposed G3 episodes that adopted the planted label; Wilson intervals. *(c)* Identity channel at G2 under each prompt (S staged, L lean), from timestamps.

prompt leaves 13, and doubles the judge's derived count from 9 to 19 under lean (Fig. 4c), so the cross-prompt comparison of provenance mixes an elicitation change with a procedure change and is reported descriptively. Read that way, grounded identity claims are 74.7% under both prompts (62/83 each; CI on the difference $[-13.1, +13.1]$) and false-premise adoption 76.1% against 77.5% (CI $[-13.2, +11.6]$). Per model, no change in adoption is detected in the three adopting families, and the Opus 5 contrast is unresolved (6/30 against 9/21, $p = 0.12$; Table 18, Fig. 4a,b). What the scaffold does change is whether an identity is committed at all.

4.4 Handling of a supplied false premise

Of the 186 exposed episodes, 141 of the 184 with a verdict adopt the false label (76.6%), a rate that mixes two populations. Three families adopt in 126/133 exposed episodes (94.7%): GPT-5.5 48/50, Sonnet 5 42/42, Gemini 36/41. Opus 5 adopts in 15/51 (29.4%), names the true cohort in 34/51 and is undecided in 2/51, so this is resistance rather than abstention (odds ratio 0.023, $p = 2.1\text{e}{-19}$). Inability to identify is not the explanation (§A.16), and the predictors measured before the premise test, reveal timing, the scaffold (§4.3) and grounded support, the preregistered predictor, are all null within every model; pooled across models the last would appear to confirm the preregistered hypothesis, by composition alone (§A.19).

Premise-test reading. Fig. 5 gives the reading of the premise test (Table 11). First, testing is common: Opus 5 tests the false label in 48 of 49 exposed episodes with a handling verdict, Sonnet 5 in 26 of 42, GPT-5.5 in 26 of 49; only Gemini, in 4 of 42, mostly computes nothing against it. Second, testing alone does not protect: in the three adopting families adoption is the same whether or not the label was tested, and Opus 5 has a single untested episode, so the comparison is unavailable there. Third, what differs is the arbitration. When Sonnet 5 tests the label it reads its own evidence as supporting it in 26 of 26 episodes, so the test never fails, although it runs the most code of any family (60 calls per exposed episode against 40 for Opus 5): the computation elaborates the supplied identity rather than testing it. When GPT-5.5's evidence contradicts the label (8 episodes) the harness prevails in 6. When Opus 5's evidence contradicts the label (42 of 51 episodes) it rejects in 28, is undecided in 6 and accepts regardless in 6 . The G0 control shows that this is calibration rather than contrarianism: with a true label supplied, Opus 5 tests it in 21 of 41 episodes with a handling verdict and Sonnet 5 in 20 of 42, and no model ever rejects a true label. Two annotators blind to the panel reproduce its reading of which side prevailed in 18/19 and 17/19 of 20 stratified traces (κ 0.83 and 0.60) but not whether the label was tested at all (15/19 and 8/19, κ 0.41 and 0.10; a third annotator was excluded as a protocol misread, §A.11), so the arbitration reading is the one the results rest on.



Figure 5: The premise test: a plausible false identity is supplied through the harness (G3) and the agent's handling of it is compared with its handling of the true identity under G0. *(a)* Share of exposed G3 episodes whose final identity is the planted label, Wilson intervals. *(b)* What the agent did with the label, upper bar the false label (G3, exposed), lower bar the true label (G0): tested, then rejected, undecided or accepted, it computed evidence bearing on the label and named a different cancer, committed to neither, or kept the label; accepted untested, it used the label with nothing computed against it; ignored, neither used nor tested; no consensus, the judges disagreed. *(c)* The same episodes, split by handling as in *(b)* and ordered by whether any result the agent computed pointed away from the supplied cancer: solid, a contradicting result existed (count at right); enclosed in the dashed box, no contradicting result was classed by the panel, whether because nothing bearing on the label was computed, the check supported it, or the judges did not agree on the direction (11 episodes). A result can contradict the label without the agent treating it as a test.

5 Limitations

Similar to other evaluation protocols, this one has several limitations. First, it is validated on one task in one domain and transfer is untested; the false premise was a plausible identity through a single channel, under a prompt that asks the agent to annotate its existing partition, so the probe also measures deference; and the process labels are judge readings of the logged trace and the naming channel a timestamp, so the protocol characterises logged evidence use, not the agent's internal mechanism. Second, the headline exception rests on one model, one task and 51 exposed episodes, and the same model recognises the shape of the benchmark, so derivation and recognition cannot be apportioned; its within-model premise-test contrast aligns with the outcome verdict by construction (§A.10), whereas the between-model difference and the zero rejection rate on true labels do not. Third, the planted condition has two cohort pairs, so per-model replication is reported in place of a clustered interval, and process levels are panel-derived: only the premise test's reading of which side prevailed is human-validated (§A.11), and one judge family is an undisclosed alias during review, every direction holding on the two public families alone (§A.6).

6 Conclusion

We described an evaluation protocol for scientific discovery agents that scores how an answer was reached separately from whether it is right. On blinded subtype discovery over TCGA cohorts it shows that withholding an identity produces derivation in a minority of episodes and recognition through channels that blinding leaves open; that the agents interpret well an analysis no better than a scripted baseline; that a prescribed procedure raises the outcome score without a detectable change in false-premise adoption; and that a supplied prior usually prevails over the agent's evidence, the exception being a model that defers to its own evidence. Overall, the results indicate that current agents are not yet reliable enough for unsupervised scientific discovery: on the one task where the answer was known, three of four frontier families recalled it or accepted a false one, and the outcome score registered neither. As these agents are placed in discovery pipelines and multi-agent systems (Gottweis et al., 2025), where no answer exists to check against, the protocol provides the means to examine an agent's reliability from its own log (§A.20), and it shows that reliability to be a property of the model rather than of the prescribed procedure, so that it must be measured before an agent is trusted rather than assumed from its score.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods (analysis, scoring and figure code, verified and tested by the authors) and to edit the paper for readability. We have not used generative AI tools to generate synthetic data, to develop the theoretical framework, to formulate or prove mathematical claims, to propose or refine hypotheses, to design the experiments, to clean or reformat datasets, to perform qualitative analysis, or to interpret results; translation and transcription are not applicable to this work. Additionally, we used generative AI tools to review parts of the manuscript, to modify, and to format references. We have reviewed all AI-assisted work: every quantitative claim was traced to the released generators and source data, and all generated code was checked by the authors. The systems studied in this paper are themselves LLM agents, and their outputs are the object of the evaluation rather than a contribution to the writing. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The study uses public, de-identified TCGA cohorts and involves no new human-subjects data; the agents never see patient identifiers. The human audit of the premise test (§A.11) was carried out by volunteer annotators from the research group, blind to model, condition and panel output, and no personal information about them is collected or reported. The protocol is intended to make evaluations of discovery agents more faithful; we see no direct path from it to harmful applications.

REPRODUCIBILITY STATEMENT

The harness, prompts, scoring rubric, judge configurations and the scripts that run and score an episode are available in an anonymized repository (`https://anonymous.4open.science/r/CancerDiscoveryEval-2EC5`). Episode logs and scoring artifacts, which the reported numbers are computed from, will be released with the camera-ready version; the harness and audits implement and check disclosure, injection, and blinding (§A.19).

REFERENCES

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023. doi: 10.1038/s41586-023-06792-0.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, 2024. doi: 10.1038/s42256-024-00832-8.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery, 2024. URL `https://arxiv.org/abs/2410.05080`.

Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15607–15631, 2023. doi: 10.18653/v1/2023.acl-long.870. arXiv:2305.01937.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *First Conference on Language Modeling (COLM)*, 2024. arXiv:2404.04475.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, et al. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

Katherine A. Hoadley, Christina Yau, Toshinori Hinoue, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304.e6, 2018. doi: 10.1016/j.cell.2018.03.022.

Shengding Hu, Yifan Luo, Huadong Wang, Xingyi Cheng, Zhiyuan Liu, and Maosong Sun. Won't get fooled again: Answering questions with false premises. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5626–5643, 2023. doi: 10.18653/v1/2023.acl-long.309. arXiv:2307.02394.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. MLAgentBench: Evaluating language agents on machine learning experimentation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, pp. 20271–20309, 2024. arXiv:2310.03302.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. DiscoveryWorld: A virtual environment for developing and evaluating automated scientific discovery agents. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. arXiv:2406.06769.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023. URL https://arxiv.org/abs/2307.13702.

Jon M. Laurent, Joseph D. Janizek, Michael Ruzo, Michaela M. Hinks, Michael J. Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D. White, and Samuel G. Rodriques. LAB-Bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, 2024.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating llms as agents, 2023a. URL https://arxiv.org/abs/2308.03688.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023b. doi: 10.18653/v1/2023.emnlp-main.153. arXiv:2303.16634.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery, 2024. URL https://arxiv.org/abs/2408.06292.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. DiscoveryBench: Towards data-driven discovery with large language models. *arXiv preprint arXiv:2407.01725*, 2024.

Ludovico Mitchener, Jon M. Laurent, Alex Andonian, Benjamin Tenmann, Siddharth Narayanan, Geemi P. Wellawatte, Andrew White, Lorenzo Sani, and Samuel G. Rodriques. Bixbench: a comprehensive benchmark for llm-based agents in computational biology. *arXiv preprint arXiv:2503.00096*, 2025.

Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52(1–2):91–118, 2003. doi: 10.1023/A:1023949509487.

Robert G. Newcombe. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine*, 17(8):873–890, 1998. doi: 10.1002/(SICI)1097-0258(19980430)17:8⟨873::AID-SIM779⟩3.0.CO;2-I.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2404.13076.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13387–13434, 2023. doi: 10.18653/v1/2023.findings-acl.847. arXiv:2212.09251.

Yusuf Roohani, Andrew Lee, Qian Huang, Jian Vora, Zachary Steinhart, Kexin Huang, Alexander Marson, Percy Liang, and Jure Leskovec. BioDiscoveryAgent: An AI agent for designing genetic perturbation experiments. *arXiv preprint arXiv:2405.17631*, 2024.

Yasin Şenbabaoğlu, George Michailidis, and Jun Z. Li. Critical limitations of consensus clustering in class discovery. *Scientific Reports*, 4:6207, 2014. doi: 10.1038/srep06207.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2023. URL https://arxiv.org/abs/2310.13548.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *PMLR*, pp. 31210–31227, 2023. arXiv:2302.00093.

E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951. doi: 10.1111/j.2517-6161.1951.tb00088.x.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Aljubeh, Mia Glaese, Ilya Sutskever, et al. PaperBench: Evaluating AI's ability to replicate AI research. *arXiv preprint arXiv:2504.01848*, 2025.

The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012. doi: 10.1038/nature11412.

The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, 2011. doi: 10.1038/nature10166.

The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–525, 2012. doi: 10.1038/nature11404.

The Cancer Genome Atlas Research Network. Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447):67–73, 2013a. doi: 10.1038/nature12113.

The Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013b. doi: 10.1038/ng.2764.

The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–550, 2014. doi: 10.1038/nature13385.

The Cancer Genome Atlas Research Network. The molecular taxonomy of primary prostate cancer. *Cell*, 163(4):1011–1025, 2015. doi: 10.1016/j.cell.2015.10.025.

The Cancer Genome Atlas Research Network. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell*, 169(7):1327–1341.e23, 2017. doi: 10.1016/j.cell.2017.05.046.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023. URL https://arxiv.org/abs/2305.04388.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*, 2023.

Matthew D. Wilkerson and D. Neil Hayes. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*, 26(12):1572–1573, 2010. doi: 10.1093/bioinformatics/btq170.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations (ICLR)*, 2024a. arXiv:2305.13300.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments, 2024b. URL https://arxiv.org/abs/2404.07972.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565, 2024. doi: 10.18653/v1/2024.emnlp-main.486. arXiv:2403.08319.

Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. Injecagent: Benchmarking indirect prompt injections in tool-integrated large language model agents. *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Yuge Zhang, Qiyang Jiang, Xingyu Han, Nan Chen, Yuqing Yang, and Kan Ren. Benchmarking data science agents, 2024. URL https://arxiv.org/abs/2402.17168.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

A Supplementary Material

Every number below regenerates from `results/tcga/clean` and `results/tcga/clean_lean` via the generator named in each subsection. Where a rate is computed over G3 episodes it is stated on the exposed denominator (episodes that were shown a planted label), for the reason given in §A.19; where we also give the arm denominator, it is labelled as the artifact. Where a contrast pools models it is labelled as pooled and accompanied by its per-model rows.

A.1 The outcome rubric

A.2 Outcome-rubric anchors

Two qualifications apply to the partition comparison in §4.2. The published partition is itself not a strong signal on the survival and structure components (0.22 and 0.11, below the agents' 0.26 and 0.21), so partition quality is measured only against the published labels and the scripted baseline; and neither interpretive component checks that a mechanism is the right one for the subtypes, only that it is evidence-backed and its pathways are real. Driver-gene coherence tells the same story as

Table 1: The outcome rubric. Six components are computed deterministically from the agent's submission; mechanism grounding is LLM-judged. The cohort-identity gate zeroes the two disease-interpretation components (†) when the agent commits to the wrong cancer.

Component	Measures	Weight
clinical signal	survival separation (ΔC-index / log-HR, Cox)	3
structure validity	bootstrap-stable silhouette + re-cluster ARI	2
genomic coherence	driver-gene enrichment	2
reference concordance	NMI between the partition and the published subtypes	2
marker evidence	discriminative power of reported markers (AUC)	2
mechanism grounding † (*LLM-judged*)	is the mechanism hypothesis evidence-backed	2
pathway validity †	pathway names valid in MSigDB/Reactome + ORA	1

NMI (0.55, against 0.54 for the k-means and 0.70 for the published partition), and 78 of 168 G0 episodes exceed their cohort's k-means NMI.

To test whether the small G0–G2 movement could be an artifact of a compressed rubric, we score three non-agent anchors with `calibrate_tcga_anchors.py`: a random partition with random genes (floor), three-cluster k-means on the top-variance genes with markers recovered from its own partition (scripted baseline), and published TCGA subtype calls (published-reference anchor). They pass through the same evaluator as the agents with the LLM calls disabled. We compare only the five components that read the submitted partition, markers and data (clinical signal, structure validity, genomic coherence, reference concordance and marker evidence), renormalised over their original weights ($11/14$ points). Mechanism grounding and pathway validity are excluded rather than entered as zeros because prose-free anchors cannot exercise them.

Table 2 makes two scopes explicit. The all-cohort floor and scripted baseline use all seven cohorts. The published reference is usable only for BRCA, OV, PRAD and UCEC; for the calibrated range, all three values are therefore recomputed on those same four cohorts. The reference is an anchor, not a literal ceiling: no biological partition is expected to optimise every internal clustering metric.

Table 2: Calibration of the five data-dependent rubric components. Values in the matched column use only the four cohorts with a sufficiently complete published subtype reference. Honest-arm values are agent submissions ($n = 501$) evaluated on the same five-component subtotal.

submission	all seven cohorts	matched four cohorts
random partition + random genes	0.118	0.134
scripted k-means	0.370	0.372
published-reference partition	—	0.492
agents, honest arms (25th/median/75th percentile)	$0.286/0.401/0.464$	—
agents by arm (G0/G1/G2)	$0.388/0.379/0.378$	—

The matched reference–floor span is 0.358, whereas the honest-arm span is 0.010. Thus the near-flat arm contrast is not explained by an inability of these five components to move. On the reference-concordance component alone (NMI against the published subtypes, for which the published partition scores 1.0 by construction), the scripted k-means reaches 0.32 averaged over the seven cohorts (BRCA 0.42, OV 0.44, LUSC 0.34, PRAD 0.32, UCEC 0.27, LIHC 0.24, LUAD 0.22) against 0.01 for the random floor; the agents' G0 episodes, with the identity disclosed, average 0.32 (§4.1), and per cohort exceed the k-means value in LIHC, PRAD and UCEC and fall below it in BRCA, LUAD, LUSC and OV (78 of 168 episodes above their cohort's baseline). On driver-gene coherence the k-means baseline averages 0.54 and the published partition 0.70; on clinical signal and structure validity the published partition scores 0.22 and 0.11, below the agents' G0 means of 0.26 and 0.21, so neither component distinguishes a good partition from a poor one on these cohorts and neither is used for that comparison. The calibration also reveals two limitations. First, the structure-validity inversion reported in §A.19 shows that internal cluster geometry is not biological correctness, the limitation of stability-based class discovery that Şenbabaoğlu et al. (2014) document for consensus clustering (Monti et al., 2003; Wilkerson & Hayes, 2010): the published partition scores below scripted k-means on that component in all four cohorts (0.16 vs. 0.45 on BRCA, 0.05 vs. 0.28

on PRAD). Second, PRAD's random-partition floor is unusually high (0.243 averaged over anchor seeds), driven by chance survival separation in a cohort with very few deaths. The anchor artifact is `manuscript/figures/rubric_anchors.json`; neither agents nor API calls were rerun.

A.3 Cluster cardinality and a k-matched baseline

NMI rises with the number of clusters, so the comparison in §4.2 is repeated here with the agents' cluster counts, adjusted mutual information (AMI), and a k-means baseline run at each episode's own k on the same anonymised matrix and seed (top-2000 variance genes, the scorer's best-scheme rule for every partition). Over the 501 honest-arm episodes with a scored partition, the pooled AMI is 0.29 against 0.31 for the fixed three-cluster baseline and 0.30 for the k-matched one; 187 episodes exceed their k-matched baseline. Cluster counts: $k = 2$ in 126 episodes, 3 in 126, 4 in 155, 5 in 71, 6 or more in 23.

Partition provenance. The partition evaluated everywhere in this paper is the one the scorer received: the submission's `proposed_grouping`, which the harness resolves from a file path when the agent passes one, falling back to the `grouping.json` the agent saved in its episode directory only when the submission is empty. In five lean-prompt episodes that saved file differs from the scored submission. In four Sonnet 5 episodes (`g0_brca_s42`, `g1_brca_s7`, `g1_ucec_s42`, `g2_ucec_s7`) the agent passed the relative path `output/grouping.json`, which resolved outside the episode directory to a 497-patient, two-cluster file from another run, so the scored partition is not the agent's own and its partition components are near zero in those four episodes. In one Gemini episode (`g1_lusc_s123`) the agent passed a valid path to a later two-cluster partition, so the saved file is an earlier draft. All five are reported as scored; rescoring the four Sonnet 5 episodes on their saved partitions would raise Sonnet 5's lean-prompt subtype recovery, and both partitions will be released with the episode logs. Four Gemini 2.5 Pro lean episodes (`g1_luad_s42`, `g2_luad_s42`, `g2_ov_s42`, `g3a_ov_mislead_brca_s7`) submitted an empty partition and have no outcome score; they remain in every process and identity analysis for which the field exists.

Table 3: Cluster counts and cardinality-corrected subtype recovery per model and arm. k: median (range) of groups committed. NMI as scored; AMI on the same scheme; AMI of the three-cluster and of the k-matched k-means baselines; episodes whose AMI exceeds their k-matched baseline.

model	arm	n	k	NMI	AMI	k=3 base	k-matched	above
GPT-5.5	G0	42	4 (2–6)	0.29	0.28	0.31	0.33	7/42
GPT-5.5	G1	42	4 (2–6)	0.27	0.26	0.31	0.32	7/42
GPT-5.5	G2	42	4 (2–5)	0.27	0.26	0.31	0.31	9/42
Sonnet 5	G0	42	4 (2–7)	0.35	0.34	0.31	0.29	20/42
Sonnet 5	G1	42	3 (2–6)	0.29	0.28	0.31	0.28	22/42
Sonnet 5	G2	42	2 (2–5)	0.24	0.23	0.31	0.27	17/42
Opus 5	G0	42	4 (3–8)	0.40	0.39	0.31	0.32	22/42
Opus 5	G1	42	4 (3–9)	0.43	0.42	0.31	0.33	25/42
Opus 5	G2	42	4 (2–9)	0.41	0.40	0.31	0.32	20/42
Gemini 2.5 Pro	G0	42	3 (2–5)	0.25	0.24	0.31	0.28	14/42
Gemini 2.5 Pro	G1	41	2 (2–5)	0.23	0.22	0.31	0.28	13/41
Gemini 2.5 Pro	G2	40	3 (2–4)	0.20	0.19	0.30	0.29	11/40

A.4 C3: support does not predict false-label adoption within any model

C3 was the preregistered primary hypothesis: an agent whose identity claim is *grounded* in computed evidence should resist a false label better than one whose claim is *non-grounded*. Table 4 gives the contrast on the exposed denominator, pooled and per model (`reveal_mechanism.py`, `explore_exploit.py`).

`PREREGISTRATION.md` §8 fixed the reading of this result in advance (report it as a null, and do not retreat to the strategy axis as a rescue), and we follow that; the strategy contrast behaves identically (pooled 50.0% vs. 89.9%; three families 88.9% vs. 95.4%, $p = 0.27$; within Opus 4/22 vs. 6/12, $p = 0.11$). The pooled row would have been reported as a confirmation had Opus 5 been

Table 4: C3. Adoption of the planted label by the support rating of the agent's own identity claim, exposed G3 episodes with a verdict. The pooled row is confounded by model and is shown only to document the artifact; the three-family and per-model rows are the test. Differences are grounded−non-grounded in percentage points with Newcombe 95% intervals (Newcombe, 1998).

stratum	grounded	non-grounded	diff (pt)	95% CI / p
pooled, four families (*confounded*)	69/104 (66.3%)	63/71 (88.7%)	−22.4	[−33.4, −9.8]
three adopting families	57/59 (96.6%)	60/65 (92.3%)	+4.3	[−4.9, +13.7]
GPT-5.5	19/20	24/25		$p = 1$
Gemini 2.5 Pro	3/4	30/34		$p = 0.45$
Sonnet 5	35/35	6/6		$p = 1$
Opus 5	12/45 (26.7%)	3/6 (50.0%)	−23.3	$p = 0.34$

added to the ladder without stratifying, which is why the generator now emits the per-model rows beside every pooled G3 contrast. The exposure correction bites hardest on the late-reveal lean cell, whose denominator falls from 64 to 8 once episodes never shown a label are removed.

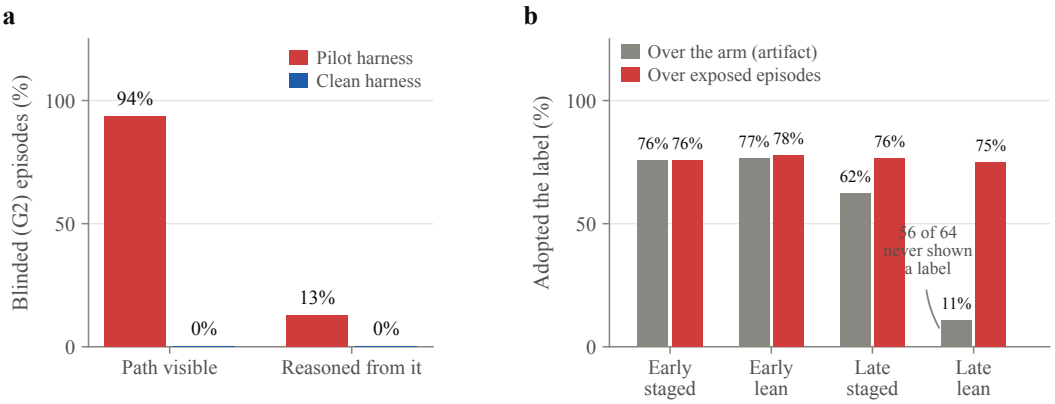


Figure 6: Two non-events recorded as benign values. *(a)* The working-directory path carried the cohort name; what corrupts a result is not that it was visible but that the agent reasoned from it, which the scrubbed harness drives to zero. *(b)* The planted label fires on the agent's Nth recorded observation with no fallback, so in the late-reveal lean cell 56 of 64 episodes were never shown a label and were scored as having resisted one. Over the arm (grey bars) that reads as a large timing effect; among exposed episodes (red bars) the rates are high and the timing contrast is not detectable.

A.5 Seed dispersion, and what it does to the G1–G2 ordering

The honest arms are a balanced three-seed design: every model × prompt × arm cell holds 21 episodes, seven cohorts under each of `s42`, `s7`, `s123`. That makes run-to-run spread directly measurable, so any contrast smaller than it is not a finding (`seed_dispersion.py`). G3 mixes seed sets over only two cohorts and is excluded here rather than pooled. At n=3 a sample SD is itself noisy, so the range is the summary we lean on; neither is a confidence interval.

Table 5: The two within-arm contrasts recomputed inside each seed, pooled over models (28 episodes per arm per seed, before no-consensus exclusions). The headline shift is positive in every seed and larger than the seed range; the G1–G2 gap changes sign.

contrast	prompt	`s42`	`s7`	`s123`	pooled	seed range
G0→G2 (headline)	staged	+45.0	+37.3	+52.2	**+45.4**	14.9
	lean	+35.1	+36.3	+36.3	**+35.6**	1.2
G1−G2 (ordering)	staged	+21.7	+22.8	**−12.2**	+10.0	35.0
	lean	**−2.1**	+12.2	**−1.1**	+3.7	14.3

The two contrasts behave differently. The headline G0→G2 shift is positive in all six prompt × seed cells, and its pooled value exceeds the seed range under both prompts, especially in the lean arm, where three seeds land within 1.2 points of each other. The G1–G2 gap does not survive: it reverses sign in one seed under the staged prompt and in two under lean, and the seed range (35.0 and 14.3 points) is larger than the pooled gap itself (10.0 and 3.7). The two withheld conditions are therefore not separable here; §4.1 makes no claim about their ordering.

Per-cell dispersion. Explore-rate seed SDs within a single model × prompt × arm cell run from 0.0 (G0 cells at the floor) to 34.2 points (GPT-5.5, lean, G1), with most G1/G2 cells between 8 and 34. Two things inflate this: a per-seed rate rests on seven episodes, and no-consensus exclusions leave 12–26 labelled episodes per pooled arm-seed. Outcome scores are far steadier over the same cells (median seed SD 0.011, median seed range 0.020), which is worth holding next to Claim B: the staged G0→G2 outcome movement (−0.005) sits inside typical seed noise, while the lean movement (−0.023) is roughly 1.1× the median seed range.

A.6 Claim A under every judge configuration

Claim A is read by an LLM panel, and §A.19 shows that a *process* label is the judge-fragile kind. Table 6 therefore recomputes the whole contrast under each single family, each two-family subset, and the full panel (`judge_sensitivity.py`). Single-family columns use that family's raw label and so have no no-consensus category; two- and three-family columns require a strict majority, which is why their denominators shrink. Intervals resample the seven cohorts and rebuild both arms inside each draw.

Table 6: The G0→G2 D2 explore-rate shift under all 14 judge configurations. The direction is positive in every one; the magnitude spans an order of magnitude. ∗ marks a cohort-clustered interval excluding zero. "n/c" counts episodes left without a label by that configuration.

prompt	judges	G0	G2	shift	cohort-clustered CI	n/c
staged	nemotron	10.7%	63.1%	+52.4	[+46.4, +60.7] ∗	0
staged	laguna	2.4%	14.3%	+11.9	[+8.3, +16.7] ∗	0
staged	qwen	3.6%	45.2%	+41.7	[+25.0, +59.5] ∗	0
staged	nemotron + laguna	0.0%	28.0%	+28.0	[+10.7, +47.8] ∗	98
staged	nemotron + qwen	0.0%	71.4%	+71.4	[+53.3, +87.5] ∗	81
staged	laguna + qwen	2.1%	28.0%	+25.9	[+14.5, +45.4] ∗	95
staged	**panel**	1.3%	46.7%	+45.4	[+29.6, +62.9] ∗	32
lean	nemotron	19.0%	63.1%	+44.0	[+33.3, +53.6] ∗	0
lean	laguna	8.3%	16.7%	+8.3	[+0.0, +17.9]	0
lean	qwen	9.5%	42.9%	+33.3	[+19.0, +48.8] ∗	0
lean	nemotron + laguna	11.6%	27.6%	+16.0	[−2.3, +38.5]	96
lean	nemotron + qwen	4.9%	64.5%	+59.6	[+30.3, +79.9] ∗	76
lean	laguna + qwen	10.0%	20.0%	+10.0	[−12.9, +31.1]	93
lean	**panel**	9.0%	44.6%	+35.6	[+17.0, +53.8] ∗	25

The direction survives; the magnitude does not. The direction survives every configuration: 14/14 positive, and 11/14 with an interval excluding zero. The magnitude does not: it ranges +8.3 to +71.4 points depending only on who judged. All three configurations that fail to separate from zero involve `laguna` under the lean prompt. That family assigns *mixed* to 295/760 episodes against 36 for nemotron and 77 for qwen (Table 9), so it depresses the explore rate in both arms and compresses the contrast rather than reversing it, which is why the sign still holds.

D1 and D3 under the same loop. "Only identity moves" rests on D1 and D3 sitting at an explore ceiling, and that ceiling is also a judged label. Under the same 14 configurations the D1 shift is +1.5 to +17.9 points, positive in all and separated from zero in 12, with G2 between 94% and 100% under every judge: the partition ceiling is real. The D3 shift is −6.7 to +12.2 points, separated from zero in only 3/14: the mechanism shift is separated from zero in only 3 of 14 configurations, but its level at G0 spans 43% to 100% across configurations. The D3 ceiling is a judge-set level, not a measured one.

This is the concrete form of the construct-validity concern raised in §A.19: had we run a single judge and happened to select `laguna`, we would have reported a lean-arm effect roughly a quarter of the size and not separable from zero, and nothing in a single-family design would have revealed the discrepancy. On that account, we report the direction as the claim and the magnitude as judge-dependent.

A.7 Card redaction: the support judge without the answer

The support judge is handed a reference card naming the true cohort and its canonical markers, as a fact-check. That gives it the answer while it grades process, so two agents reasoning identically could receive different provenance labels because one was right. We reran the same three judge families on the same distilled traces with the card removed and the instruction not to guess the cancer type (`score_support.py --no-card`; 760/760/759 episodes, the one gap a Gemini mislead episode on which `laguna` never returns a complete verdict), and compared labels episode by episode (`card_redaction.py`).

Table 7: Label stability when the card is removed. Agreement and Cohen's κ between the with-card and no-card labels of the same judge on the same episode; the panel row compares the two strict-majority labels where both exist.

	D2 strategy			D2 support		
judge	agree	κ	n	agree	κ	n
nemotron	64.6%	0.34	760	84.6%	0.65	760
laguna	48.1%	0.16	759	76.2%	0.46	759
qwen	70.0%	0.43	760	85.5%	0.67	760
panel	71.6%	0.46	536	88.3%	0.72	741

Table 8: The measurements that rest on the support judge, with and without the card. Claim A is the G0→G2 explore-rate shift on the panel majority; the last column is a cohort-clustered 95% interval on (no-card shift − card shift). C3 gives adopted/n for grounded vs non-grounded claims among exposed G3 episodes with a verdict.

		with card	without card	difference CI
Claim A shift	staged	$1/76 \to 28/60$ (+45.4)	$1/78 \to 46/71$ (+63.5)	$[-2.2, +38.0]$
	lean	$7/78 \to 29/65$ (+35.6)	$9/80 \to 42/70$ (+48.8)	$[+2.6, +22.2]$
D2 grounded	G0	80.1%	73.8%	
	G1	77.1%	77.4%	
	G2	74.7%	79.8%	
C3, adopted	GPT-5.5	19/20 vs 24/25	36/36 vs 12/14	
	Gemini	3/4 vs 30/34	10/11 vs 25/29	
	Opus 5	12/45 vs 3/6	13/41 vs 1/7	
	Sonnet 5	35/35 vs 6/6	35/35 vs 7/7	

The card did not produce the shift: without it the panel tags more identity decisions explore (254 vs 197 of 760), reaches consensus more often (106 vs 153 no-consensus), and the G0→G2 shift is larger, detectably so under the lean prompt and on the two public judges alone (staged $+71.4 \to +95.0$, CI $[+9.1, +41.1]$). Claim A as reported is therefore conservative. The support label is substantially stable (κ 0.65–0.67 for the public judges), so the 75.6% grounding figure and every within-model C3 contrast are unaffected. And the strategy label is only moderately stable (κ 0.34–0.43; 0.16 for `laguna`, whose habit of tagging *mixed* is the source), which is the same judge fragility §A.6 documents and the reason the paper claims direction rather than magnitude for that label.

A.8 Judge reliability and the three-family panel

Provenance, support, the CoT summary and the premise test are read by one pass each from three judge families (`nemotron-3-super`, `laguna`, `Qwen3.6-27B`). The design estimates cross-

family robustness; it does not separate family bias from per-judge stochasticity, which would need a second pass from one family. We therefore never claim that a delta exceeds judge noise.

Outcome-gate verdicts are near-invariant. On the 256 G3 episodes, the fraction the gate calls fooled is $142/256$ (55.5%) for nemotron, $139/256$ (54.3%) for laguna and $142/256$ (55.5%) for qwen, a spread of 1.2 points. Claim C does not depend on which family grades it. On the continuous outcome score the mean per-episode spread across families is 0.010 (max 0.207), and $112/760$ episodes have at least one family disagreeing about the identity gate.

Process labels are not invariant. On the identical episodes, the D2 strategy tag splits as in Table 9: the *explore* count ranges from 102 to 401, a fourfold spread, driven largely by how readily each family uses the *mixed* category. $153/760$ episodes reach no majority and are excluded from every strategy rate rather than folded into one. On `identity_derivation` the three families are unanimous on only $400/760$ episodes (52.6%), and the data-derived count runs 564 / 642 / 549. Three passes of a single model could not reveal this, because a model largely agrees with itself.

Table 9: Per-family D2 identity strategy labels, all 760 episodes, one pass each.

family	explore	exploit	mixed	`identity_derivation`=data-derived
nemotron-3-super	401	323	36	564
laguna	102	363	295	642
Qwen3.6-27B	222	461	77	549

No-consensus counts. Reported per field, never folded: D2 strategy $153/760$; D3 mechanism strategy 39; `identity_derivation` 32; D2 support 15; D3 support 13; D1 strategy 8; the identity gate 2; D1 support 1.

Gemini's traces are the hardest to read. The panel returns `not-established` for the identity derivation on $19/190$ Gemini episodes (10.0%) versus $2/190$ for GPT-5.5 and $0/190$ for Sonnet 5 and Opus 5, with 16, 7, 8 and 1 no-consensus episodes respectively. This is a statement about the readability of the trace as much as about the agent, and is why Gemini is reported separately in every cross-model process comparison.

A.9 Two provenance constructs, and why one score cannot carry both

Provenance is read by two instruments. The *support* judge tags the D2 identity decision *explore*/*exploit* and rates its support; the *CoT* judge assigns `identity_derivation`. Both are naturally described as "did the agent derive the identity", and on 259 of the 445 episodes in the binary cross-tab below they disagree. The disagreement is not noise, and its structure is itself a result.

Table 10: Identity provenance (support judge) against analytic provenance (CoT judge), over honest and exposed-G3 episodes with binary majority labels from both instruments. Mixed, not-established, no-consensus and unexposed-G3 episodes are excluded. The near-empty top-right cell is the expected direction: an identity cannot be earned from data without relevant analysis.

	analysis computed (`data-derived`)	analysis recalled (`recalled-prior`)
identity *earned* (explore)	149	**1**
identity *supplied* (exploit)	**258**	37

The constructs nest. Earning the cohort identity requires having done the analysis, but doing the analysis does not require earning the identity. Identity provenance is therefore a strict subset of analytic provenance, and the reverse cell is nearly empty ($1/445$). The strongly directional 258:1 off-diagonal is consistent with nested rather than interchangeable constructs. The CoT judge need not be malfunctioning: its label can coherently describe the broader analysis while failing to measure where the identity came from.

G0 is the calibration arm that separates them. At G0 the cohort is named in the prompt, so the identity cannot have been earned except by independent re-derivation, and the correct rate is known in advance to be near zero. The support judge returns $8/154$. `identity_derivation` returns $115/148$ because agents at G0 genuinely are computing, even though the arm does not establish whether every such analytic-provenance label is correct. Per arm, the fraction of *exploit*-tagged episodes the CoT judge nonetheless calls data-derived is $89/118$ (G0), $43/43$ (G1), $51/51$ (G2) and $75/83$ (exposed G3).

The off-diagonal cell measures the mechanism of interest. The 258-episode cell is the population this paper is concerned with: real computation organised around a premise the agent did not earn. On the exposed G3 set it contains 75 episodes, 67 of which adopted the false label. Their mean outcome is 0.227 against 0.289 for the other 44 exposed episodes carrying binary labels from both instruments; the gap is the identity gate, since the second group now holds most of Opus's resisting episodes, whose disease components are not zeroed. The sharpest members are 22 episodes rated *anchored* by the support judge (the identity is anchored to the supplied label) and `data-derived` by the CoT judge on the same trace; 19 adopted the label.

In a LUSC cohort mislabelled LUAD, one agent built a stable expression partition from opaque genes, then read those clusters as LUAD programs while its own marker output carried strong squamous signal (`SOX2`, `TP63`, `KRT5`, `KRT6A`, `NFE2L2`). Every support judge called the identity exploit/anchored; every CoT judge called the surrounding analysis data-derived. Under the two-construct interpretation, those labels describe different aspects of the trace: the partition was derived; the identity was supplied.

Implications for practice. No post-processing recovers identity provenance from a measurement of analytic provenance: the information is not in the labels, and any rule that forces the two into agreement is the support judge under another name. What is available instead is to name the constructs separately, report each from its own instrument, and include a known-source arm where the identity-provenance value is fixed a priori. This paper uses G0 for that check: every identity-provenance claim is made from the support instrument, and `identity_derivation` appears only as a descriptor. The human audit below is still required to estimate the accuracy of both instruments.

Status. The premise test has been audited by blinded annotators on 20 stratified exposed traces (§A.11). The two provenance instruments have not: a 120-episode blinded package (representative and challenge splits) is built and not yet annotated, so neither instrument has a human-validated accuracy, and we make no population claim about why each episode accepted the label beyond what the premise test records.

A.10 The premise test

Design. For every G0 episode and every G3 episode, a judge reads the same distilled trace used by the CoT summary (per-step intent, hypothesis checkpoints, final submission) with one line added: what cohort identity the harness supplied and at which observation. In G0 that is the true identity, stated in the task preamble; in G3 it is the planted one, injected at the third or fifth observation. The judge does not see the true cohort, the arm, or the outcome verdict. G3 episodes that never received the label are recorded as unexposed without a judge call. The judge returns four fields: *handling* (tested and rejected, tested and undecided, tested and accepted, accepted untested, ignored), the *direction* of the agent's own evidence (supports, contradicts, mixed, none computed), the *final relation* of the submitted identity to the supplied one, and a short quoted *evidence* span. Three families judge every episode; each field is reduced by strict majority (`score_premise_test.py`, `premise_test.py`). Because `panel_status.py` does not know this artifact, the analysis refuses to run on an incomplete lane.

Reading. Testing alone protects nothing: within every model, adoption is the same among tested and untested episodes (GPT-5.5 $25/26$ vs. $21/22$; Sonnet 5 $26/26$ vs. $16/16$; Gemini $4/4$ vs. $32/37$). What separates the models is what happens when the test fails (Fig. 7). Sonnet's tests never fail: in all 26 tested episodes it reads its evidence as supporting the planted label, which is the retrofit of §A.9 seen from the agent's side. GPT-5.5's evidence contradicts the label in 8 episodes and the

Table 11: Premise test, exposed G3 episodes, panel majority per field, with the per-family "tested" rate and the G0 true-label control. Tested rates are over episodes with a handling verdict (49, 42, 42, 49 exposed; 42, 42, 42, 41 at G0).

	GPT-5.5	Sonnet 5	Gemini 2.5 Pro	Opus 5
exposed G3, n	51	42	42	51
tested (panel majority)	53.1%	61.9%	9.5%	98.0%
per family (laguna/nemotron/qwen), %	41/63/53	48/64/55	7/19/14	96/98/96
tested → rejected/undecided/accepted	1/1/24	0/0/26	0/0/4	28/6/14
untested/ignored/no consensus	23/0/2	16/0/0	36/2/0	1/0/2
evidence contradicted the label, n	8	2	0	42
then rejected/undecided/accepted/untested	1/0/6/0	0/0/0/2	—	28/6/6/0
adopted (outcome gate)	7/8	2/2	—	8/42
tested, accepted, evidence read as supporting	14	26	4	4
G0, n	42	42	42	42
tested	19.0%	47.6%	14.3%	51.2%
rejected a true label	0	0	0	0

harness wins in 6. Opus's evidence contradicts the label in 42 episodes and the evidence wins in 34. Gemini computes nothing against the label in 36 of 42 episodes.

Consistency and circularity. The judge's blind *final relation* agrees with the outcome gate's verdict in 167/182 exposed episodes (92%), so the judge is reading the traces correctly. That agreement is a check on the instrument, not a result: the final relation and the verdict are the same event read twice. What is not by construction is the *handling* field, which is assigned before the verdict is known and differs between models that reach the same verdict, and the G0 control, in which a model that tests true labels accepts all of them. The within-model contrast between rejecting and accepting episodes is, by construction, aligned with adoption; we do not quote it as a test.

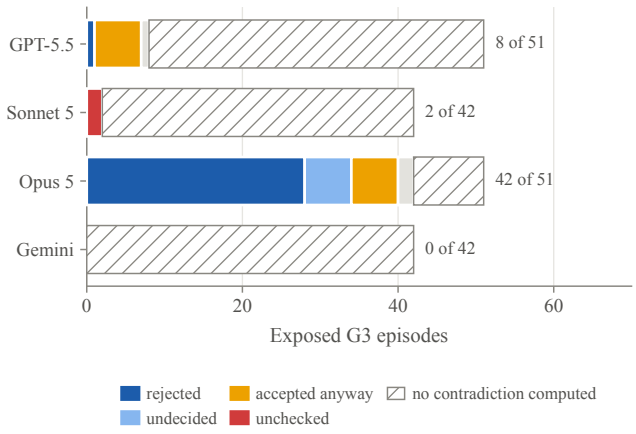


Figure 7: Every exposed G3 episode per model. Coloured segments are episodes in which the agent's own computed evidence contradicted the planted label (panel majority), split by what the agent did next; the hatched remainder was not classed as contradicting (nothing computed, supporting, or no consensus on direction). Gemini's bar is entirely hatched: an agent that rarely tests the label (4 of 42) rarely reaches a contradiction to arbitrate.

Representative episodes. Three episodes illustrate the readings. In a LUSC cohort labelled LUAD, one agent built a stable expression partition from the opaque genes and, once the label arrived, read the clusters as LUAD programs while its own marker output carried a strong squamous signal (SOX2, TP63, KRT5, KRT6A, NFE2L2); every support judge rated the identity anchored and every analytic-provenance judge rated the surrounding analysis data-derived (§A.9). This is the population of 109 episodes with an explore-tagged mechanism for the wrong cancer: the computation

is real and is organised around a premise the agent did not earn. In a second episode, GPT-5.5 had inferred LUSC before the label arrived, wrote "my earlier inference of LUSC is contradicted by the source-cohort reveal," and submitted LUAD; the evidence was computed, compared and overruled. In a third, an OV cohort labelled BRCA, Opus 5 checked the breast-defining alterations (PIK3CA, CDH1, GATA3) and luminal markers and found them absent, confirmed the ovarian markers (PAX8, WT1, MSLN, WFDC2) and TP53 mutation in 94% of samples, wrote "the provided source tag says TCGA BRCA, but the molecular evidence contradicts a breast cohort," and submitted high-grade serous ovarian carcinoma. The behaviour of Opus 5, exploiting the prior when the evidence agrees and exploring when it does not, is the balance the protocol is built to detect; on one model, one task and two cohort pairs, it was observed.

A.11 Human audit of the premise test

Design. Twenty exposed G3 episodes were drawn to cover the panel's handling categories rather than the population: two accepted untested, five tested with supporting evidence (all Sonnet 5), nine tested against contradicting evidence and accepted anyway (six GPT-5.5, three Opus 5), two rejected and two undecided (all Opus 5). Each case page showed the supplied label, the agent's stated intent and hypothesis checkpoints after it, and the submission, with cancer and tissue names highlighted; annotators did not see the model, prompt, arm, cohort or any panel output. They chose one letter per case: A, did not check the label, used it; B, checked, own results fit, used it; C, checked, results did not fit, used it anyway; D, checked, did not fit, said the cancer was something else; E, checked, did not fit, did not commit; F, ignored the label; U, cannot tell. "Checked" was defined as having run an analysis on the dataset and compared its result with the label; repeating the label or naming clusters after it does not count. The letters map onto the panel's handling categories, B and C both corresponding to tested-accepted and distinguished by the direction of the evidence (Table 12). Three annotators returned sheets. One marked 18 of 20 cases E, one B and one U, with notes listing the submitted cluster names as evidence that no cancer was committed; that reading treats subtype labels as the identity commitment, contradicts the other two annotators on every case ($0/20$ agreement with either), and is excluded as a protocol misread. Agreement is computed with `premise_audit.py` and reported per distinction, with U retained as its own class for the six-way letter.

Table 12: Human audit of the premise test. Per panel category, the letters assigned by the two retained annotators; below, agreement with the panel and between annotators on the six-way letter, on which side prevailed (A/B/C versus D/E) and on whether the label was tested (A versus B–E), with Wilson 95% intervals and Cohen's κ.

panel category (letter)	n	annotator 1	annotator 2
accepted untested (A)	2	A 2	A 2
tested, supports, accepted (B)	5	B 2, C 2, U 1	A 5
tested, contradicts, accepted (C)	9	C 5, A 4	A 6, C 1, B 1, D 1
tested, rejected (D)	2	D 2	D 2
tested, undecided (E)	2	D 1, C 1	C 1, U 1
agreement with panel, six-way letter		$11/20$, κ 0.40	$5/20$, κ 0.13
which side prevailed		$18/19$ (75–99%), κ 0.83	$17/19$ (69–97%), κ 0.60
tested versus untested		$15/19$ (57–91%), κ 0.41	$8/19$ (23–64%), κ 0.10
between annotators, six-way letter		$10/20$, κ 0.32	
which side prevailed		$17/18$, κ 0.77	
tested versus untested		$12/18$, κ 0.40	

Reading. The two distinctions behave differently. Which side prevailed is reproduced by both annotators in all but one and two cases respectively, and by the two annotators with each other in 17 of 18; every rejection and every untested acceptance is reproduced by both; of the seven Opus 5 cases, one undecided case was read as accepted by both annotators, crossing the boundary. Whether the label was tested is not reproduced: the second annotator applied the definition strictly and counted an explicit comparison of computed results with the label as the only test, under which all five Sonnet 5 "supports" cases and six of nine "contradicts" cases become untested acceptances. The panel's tested rate therefore includes episodes in which the agent re-interprets results it has

already computed under the new label. The two undecided cases were resolved by the annotators to a rejection or an acceptance, so the undecided category is not reproduced at this sample size. The sample is stratified, not representative, so these are agreement rates per category and not accuracy estimates for the benchmark population.

A.12 How a withheld identity was decided

Table 13 joins the support judge's recall type (panel majority) to the first point in the trace at which the agent names a disease, placed relative to the codebook reveal (`identity_channel.py`). "First named" reads the agent's stated intent and hypothesis checkpoints, not its code or printed output, so it is an upper bound on how late the identity was decided. Pre-reveal namings are classified by the cues present in the naming text: dataset shape (sample count, death count, age, stage missingness, gene count) and statistics computed on opaque genes (mutation frequency, mutual exclusivity, hypermutation, copy-number burden).

Table 13: G2 identity decisions, 168 episodes. The exploit-tagged half comprises three different behaviours, and pre-reveal identification is almost entirely confined to one model.

	GPT-5.5	Sonnet 5	Gemini 2.5 Pro	Opus 5
strategy: explore/exploit/mixed/n.c.	11/17/4/10	18/5/6/13	14/25/0/3	14/9/2/17
recall: derived/grounded recall/bare/none	7/29/1/3	12/29/0/1	0/3/6/31	9/33/0/0
first named: pre/post/final only/never	0/36/0/6	6/35/0/1	0/6/5/31	24/18/0/0
pre-reveal cues: shape+bio/bio/shape/uncl.	—	2/3/1/0	—	15/5/1/3

Pooled, the 56 exploit-tagged episodes are 30 grounded recalls (named after the reveal, then confirmed), 17 non-identifications, 7 bare assertions and 2 without consensus; 30 episodes name a disease before any gene symbol is visible and 29 name the true cohort. The sample-count probe of §A.19 is the strict subset of these in which the exact sample size sits beside the naming; the fuller cue set above is why Opus's pre-reveal rate ($24/42$) exceeds the probe's ($10/42$).

A.13 Self-reported confidence (descriptive only)

The agent's first stated confidence falls from G0 to G2 in $8/8$ model $\times$ prompt lanes (GPT staged $0.46 \to 0.37$, GPT lean $0.31 \to 0.25$; Sonnet staged $0.38 \to 0.32$, Sonnet lean $0.49 \to 0.40$; Opus staged $0.58 \to 0.54$, Opus lean $0.40 \to 0.36$; Gemini staged $0.74 \to 0.71$, Gemini lean $0.49 \to 0.45$). That is a response to the disclosure arm. It is not a read on strategy: within a fixed arm, explore and exploit episodes report almost the same confidence ($+0.014$ at G1, -0.030 at G2; pooled over the honest arms 0.484 vs. 0.494, Mann–Whitney $p = 0.62$). Confidence therefore cannot be used as evidence that an answer was derived rather than recalled, and we do not use it that way. Values are the first numeric confidence in the observation log, available for $489/504$ honest-arm episodes.

A.14 Cohort heterogeneity, per cohort

Table 14 gives the blinded (G2) arm by cohort. Outcome ranges 0.363 (OV) to 0.537 (UCEC), a $1.48\times$ spread. We pair the scaffold ablation within cohort and use cohort-clustered intervals, but do not infer that lower outcome scores measure cohort difficulty or cause greater exploration.

Table 14: The blinded (G2) arm by cohort, pooled over models and prompts (The Cancer Genome Atlas Network, 2012; The Cancer Genome Atlas Research Network, 2017; 2014; 2012; 2011; 2015; 2013a). Explore and grounded denominators differ because they exclude different no-consensus episodes.

	BRCA	LIHC	LUAD	LUSC	OV	PRAD	UCEC
episodes	24	24	24	24	24	24	24
outcome	0.517	0.490	0.523	0.402	**0.363**	0.471	**0.537**
D2 explore	6/17	9/17	11/20	13/18	11/19	3/16	4/18
D2 grounded	20/24	18/23	18/24	17/23	15/24	18/24	18/24

A.15 COST

Token counts are measured per turn from `run_log.usage_log`; dollar figures are derived from a price table and should be re-checked against current provider pricing before being quoted (`cost_ledger.py`). The 760 episodes consumed 645.9M input and 22.8M output tokens (an input:output ratio of 28.4:1) for $2,515 total: Opus 5 $1,253, Sonnet 5 $678, GPT-5.5 $475, Gemini 2.5 Pro $110.

The dominant cost term is conversation length rather than answer length. Every turn re-sends the whole transcript, so cost scales with turns^2; the worst lane (Sonnet lean, 53.3 turns/episode) sends 1.44M input tokens per episode. This also explains why the lean prompt is so much cheaper for some models and not others, since it changes turn count rather than price: GPT-5.5 $\$318 \to \$157$ (-51%, $16.3 \to 12.4$ turns), Gemini $\$70 \to \$40$ (-43%, $24.5 \to 19.1$), Opus 5 $\$749 \to \$503$ (-33%, $38.5 \to 36.9$), Sonnet 5 $\$363 \to \$314$ (-13%, $52.3 \to 53.3$ turns, essentially unchanged). Normalised per 10M tokens the models are $60.84 / $56.57 / $21.81 / $15.60 (staged) for GPT / Opus / Sonnet / Gemini.

Judging is cheap relative to running. The CoT-summary judge alone made 2,280 calls totalling 17.4M tokens, and the premise test 1,062 calls totalling 7.6M (1,272 artifacts, 210 of them unexposed placeholders with no call). They cost this project nothing, because all three families run on infrastructure already available here, so the judge-to-agent cost ratio is 0% by construction and is not a number to generalise from. On the same measured token counts, a commodity hosted endpoint at $0.30/M in and $0.60/M out would charge $6.71 for the CoT summaries, which is 0.27% of the agent spend and $0.0088 per episode scored. This is the figure a reader who pays for judging should use.

A.16 IDENTITY ACCURACY AND THE G3 UNAIDED BASELINE

Table 15 separates the submitted identity from its provenance label. Accuracy is the panel's `true_cohort` verdict over episodes with a majority identity verdict; undecided answers are retained as their own category. G3 rows include only episodes that received the planted label. The two exposed G3 rows each omit one no-consensus verdict.

Table 15: Identity verdicts across disclosure arms and on the two G3 source cohorts. "Planted" is the G3 false-label verdict; it is not applicable to honest arms.

condition	true cohort	undecided	other	planted
G0 (all cohorts)	161/168	7/168	0/168	—
G1 (all cohorts)	137/168	24/168	7/168	—
G2 (all cohorts)	136/168	27/168	5/168	—
G2 OV unaided	15/24	8/24	1/24	—
G3 OV→BRCA	18/90	6/90	0/90	66/90
G2 LUSC unaided	19/24	4/24	1/24	—
G3 LUSC→LUAD	18/94	1/94	0/94	75/94

Pooled over OV and LUSC, unaided G2 accuracy is $34/48$ (70.8%), versus $36/184$ (19.6%) after G3 exposure. The independent-proportions descriptive difference is 51.3 points (95% CI $[35.9, 63.3]$); because G2 and G3 use unequal seed sets and only two source cohorts, we lead on the two cohort rows rather than treating that interval as a clustered causal estimate. Of the 14 unaided errors, $0/9$ OV episodes already name BRCA and $1/5$ LUSC episodes already name LUAD. Within the honest arms, explore is not detectably more accurate than exploit at G1 (75.8% vs. 70.8%, CI $[-11.2, +21.6]$) or G2 (77.2% vs. 69.6%, CI $[-8.7, +23.3]$). Per model under G2: Opus 5 $42/42$ correct, Sonnet 5 $40/42$, GPT-5.5 $39/42$, Gemini 2.5 Pro $15/42$ with 23 undecided. All values regenerate from `identity_accuracy.py` and `figures/identity_accuracy.json`.

A.17 FULL COMPONENT AND PER-MODEL TABLES

Scaffold, per component. Table 17 is the data behind Fig. 8. Deltas are staged−lean over the 249 paired honest-arm episodes. Three points about how they are computed. They read un-gated

Table 16: Per-model summary of the four instruments. Disclosure ladder: D2 explore rate (%) at G0→G2 under each prompt; at G2, the point at which the identity was first named relative to the codebook reveal (before / after / in the final submission only / never, of 42 episodes) and the number of correct identity verdicts. Outcome components at G0: subtype recovery (NMI against the published subtypes; the scripted k-means baseline is 0.32) and mechanism grounding, both on the 0–1 raw scale. Scaffold: paired staged−lean outcome delta. False premise, on exposed G3 episodes: planted label adopted (over episodes with a verdict); share in which the label was tested; among episodes whose own evidence contradicted the label, the count and how many were rejected / undecided / accepted (accepted includes accepted untested; the remainder had no consensus on handling). True premise, on G0: share that tested the true label and the number that rejected it. Detail in Tables 19, 13, 15, 18 and 11.

instrument	reading	GPT-5.5	Sonnet 5	Opus 5	Gemini 2.5 Pro
disclosure ladder	explore G0→G2 (staged)	5.0 → 37.5	0.0 → 58.3	0.0 → 53.8	0.0 → 42.1
	explore G0→G2 (lean)	5.0 → 31.2	0.0 → 64.7	5.6 → 58.3	25.0 → 30.0
	G2 named pre/post/final/never	0/36/0/6	6/35/0/1	24/18/0/0	0/6/5/31
	G2 identity correct	39/42	40/42	42/42	15/42
outcome (G0)	subtype recovery, NMI	0.29	0.35	0.40	0.25
	mechanism grounding	0.99	0.99	1.00	0.75
scaffold	Δ outcome, staged−lean	+0.000	+0.073	+0.037	+0.064
false premise (G3)	planted label adopted	48/50	42/42	15/51	36/41
	label tested	53%	62%	98%	10%
	contradicted: n, rej/undec/acc	8: 1/0/6	2: 0/0/2	42: 28/6/6	0
true premise (G0)	label tested / rejected	19% / 0	48% / 0	51% / 0	14% / 0

scores: the evaluator writes `raw_scores` before the cohort-identity gate and the gate zeroes only `weighted_scores`, so no component here is suppressed by a fooled-identity verdict, and in any case the pairs are honest arms, where no false label exists. Seven components are tested on one dataset, so we report Holm-adjusted p beside the raw Wilcoxon. And the 249 pairs come from only seven cohorts, so the cohort-clustered bootstrap interval (the same 7-cluster resampling used for the pooled delta) is the honest precision statement.

Table 17: Staged−lean per rubric component, paired and un-gated. Holm is applied across the seven-component family; intervals resample the seven cohorts. Only the four marked ✓ survive both, and those are the only ones we claim.

component	Δ	Wilcoxon p	Holm p	cohort-clustered 95% CI	
pathway validity	+0.269	3.5e−33	2.5e−32	[+0.246, +0.295]	✓
reference concordance	+0.092	8.7e−12	5.2e−11	[+0.038, +0.149]	✓
genomic coherence	+0.089	3.3e−5	1.3e−4	[+0.026, +0.153]	✓
structure validity	+0.035	4.7e−5	1.4e−4	[−0.004, +0.074]	
marker evidence	+0.009	0.033	0.040	[−0.016, +0.032]	
clinical signal	−0.033	0.020	0.040	[−0.072, +0.000]	
mechanism grounding	−0.032	1.4e−6	7.2e−6	[−0.042, −0.023]	✓

The Holm correction changes nothing: all seven components are significant at $p < 0.05$ before and after. Cohort clustering changes three things: structure validity, marker evidence and clinical signal have intervals that include zero, so their apparent movements are not separable from cohort-to-cohort variation and we do not claim them. Structure validity is independently suspect: the published subtype partition scores lower on it than scripted k-means in all four cohorts where a ceiling exists, because the clustering geometry it rewards is what k-means optimises and real biological subtypes do not.

Scaffold, per stratum. Paired staged−lean outcome: all honest arms +0.043 ($n = 249$, $p = 1.3\text{e}{-13}$, staged wins 183/249); by arm G0 +0.029 ($p = 6\text{e}{-4}$), G1 +0.052 ($p = 6.5\text{e}{-7}$), G2 +0.050 ($p = 1.3\text{e}{-5}$); by model Sonnet 5 +0.073 ($p = 1.8\text{e}{-7}$), Gemini +0.064 ($p = 4.0\text{e}{-4}$),

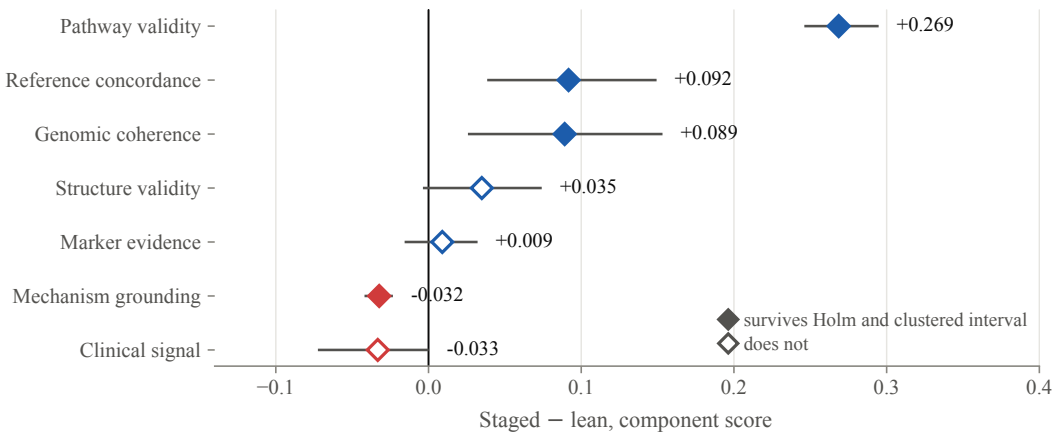


Figure 8: The scaffold redistributes component scores rather than lifting them uniformly. Staged−lean per component on un-gated raw scores, $n = 249$ pairs, with cohort-clustered 95% intervals; filled diamonds survive both Holm across the seven components and a clustered interval excluding zero, open diamonds do not.

Opus 5 $+0.037$ ($p = 2.5\text{e}{-7}$), GPT-5.5 $+0.000$ ($p = 0.99$). The cohort-clustered interval on the pooled estimate is $[+0.025, +0.062]$ over 7 cohorts and excludes zero.

Scaffold, per model (rule E4). Table 18 shows the scaffold contrast within each family before it is pooled. Grounded identity and documented rigour are read on the 21 paired G2 twins per model (Gemini: 19 with an outcome on both sides); adoption is read on exposed G3 episodes.

Table 18: Staged versus lean within model. Outcome is the paired delta (Wilcoxon p); the G2 columns give grounded / labelled twins and high-rigour / labelled twins, staged then lean; adoption is adopted / with-verdict on exposed G3 episodes, staged versus lean (Fisher p).

model	outcome Δ (p)	G2 grounded	G2 rigour high	G3 adoption, staged vs. lean
GPT-5.5	$+0.000$ (0.99)	18/21, 20/21	21/21, 17/21	30/31 vs. 18/19 (1.00)
Sonnet 5	$+0.073$ (1.8e−7)	20/21, 21/21	21/21, 18/21	26/26 vs. 16/16 (1.00)
Opus 5	$+0.037$ (2.5e−7)	21/21, 21/21	21/21, 21/21	6/30 vs. 9/21 (0.12)
Gemini 2.5 Pro	$+0.064$ (4.0e−4)	3/20, 0/20	13/21, 4/21	24/26 vs. 12/15 (0.34)

The ladder, per model. Table 19 and Fig. 9 give the explore-tagged rates behind the manipulation check in §4.1. The D2 explore rate rises from G0 to G2 in all eight lanes, but the sequence is non-monotonic: G1 exceeds G2 in five. The outcome does not move consistently in either direction.

Table 19: D2 explore rate (%) and mean outcome across the disclosure conditions, per model × prompt. $n = 21$ episodes per cell.

		D2 explore (%)			outcome		
model	prompt	G0	G1	G2	G0	G1	G2
GPT-5.5	staged	5.0	60.0	37.5	0.483	0.489	0.495
	lean	5.0	25.0	31.2	0.496	0.487	0.482
Sonnet 5	staged	0.0	61.5	58.3	0.535	0.530	0.494
	lean	0.0	73.3	64.7	0.473	0.422	0.444
Opus 5	staged	0.0	35.7	53.8	0.532	0.536	0.531
	lean	5.6	58.3	58.3	0.487	0.497	0.505
Gemini 2.5 Pro	staged	0.0	66.7	42.1	0.453	0.437	0.461
	lean	25.0	36.8	30.0	0.432	0.373	0.354

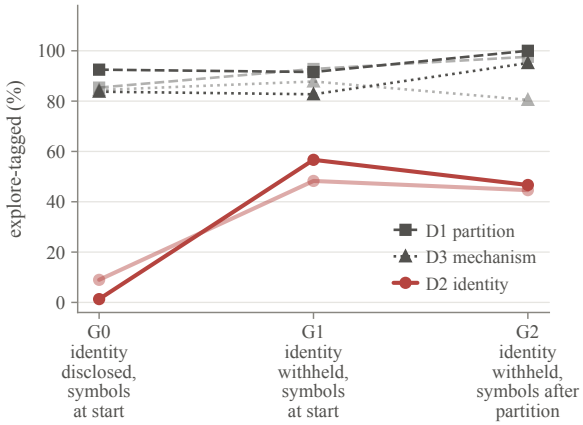


Figure 9: Explore-tagged share per commitment across the ladder (panel majority; dark staged, light lean; line style marks the commitment). Partition and mechanism sit at their ceilings in every arm; only identity moves, and that movement is the manipulation check, not the finding.

Adoption, per model (exposed). GPT-5.5 $48/50$ (96.0%), Sonnet 5 $42/42$ (100%), Gemini 2.5 Pro $36/41$ (87.8%), Opus 5 $15/51$ (29.4%); pooled $141/184$ (76.6%), three adopting families $126/133$ (94.7%). By cohort pair: LUSC→LUAD 79.8%, OV→BRCA 73.3%.

A.18 Prompts and the disclosure mechanics

Both prompts are included in the reproducibility package: the staged prompt (304 lines) and lean prompt (83 lines), at `prompts/agent_system_tcga.txt` and `prompts/ablation/tcga_lean.txt`. They are identical in task, tools, budget and codebook rules, and differ in whether a procedure is prescribed and, as §4.3 notes, in whether the cancer identity is elicited.

System identification. The agent model identifiers recorded in every episode are OpenAI `gpt-5.5`, Anthropic `claude-sonnet-5` and `claude-opus-5`, and Google `gemini-2.5-pro`. Calls use a common 32k per-turn output budget, request no temperature or sampling seed, and leave provider reasoning at its deployed default; no explicit extended-thinking budget is requested. The judge panel is `nemotron-3-super` (NVIDIA Nemotron), `laguna` (an internal alias whose underlying checkpoint will be disclosed after double-blind review), and `Qwen/Qwen3.6-27B-FP8`, one pass per family through institution-hosted OpenAI-compatible endpoints. The alias prevents independent reconstruction of its sensitivity rows during review; we therefore report that limitation rather than implying public reproducibility. Per-artifact judge provenance is included with the package; "one pass per family" measures cross-family sensitivity, not within-model stochasticity.

Staged. Seven numbered stages the agent is told to work through in order: Stage 0 data orientation, 1 signal discovery, 2 subtype inference, 3 analysis plan, 4 statistical deep-dive, 5 biological annotation and validation, 6 final report and submission. Each stage names the analyses expected of it.

Lean. The same task statement with the procedure removed and its absence made explicit: *"There is NO prescribed procedure. Choose your own methods and analyses—decide what to compute, in what order, and how to validate it. You are the scientist; the approach is yours to design."*

What both share. The tool budget and forced-submission point; the instruction that `submit_discovery` ends the session; and the checkpoint block governing gene identity. That block carries the codebook rule in both prompts: genes are opaque `GENE_XXXXX` identifiers, the codebook arrives on the agent's third call to `record_observation`, and when it arrives the agent is told to *"annotate the clusters you already found—do NOT rebuild the partition from biological*

priors." Because the rule is in both, the scaffold ablation isolates the staged procedure and not the codebook discipline.

The ladder and the mislead injection. The disclosure level is set by a single `disease_hint` slot in the shared preamble: G0 states the cohort, G1 leaves it empty while releasing real gene names, and G2 leaves it empty under full anonymisation. G3 builds on the G2 base and injects a false cohort label through the same codebook-reveal channel, at the agent's third recorded observation (G3a) or fifth (G3b). The injection has no call-count fallback, which is the defect analysed in §A.19: an episode that submits before its Nth observation is never shown a label at all, and must be excluded rather than scored as having resisted one.

The factorial design assigns 128 episodes to each reveal time (2 cohort pairs $\times$ 4 models $\times$ 8 seeds $\times$ 2 prompts). Conditioning on intervention delivery leaves 125 early and 59 late exposed episodes with a verdict (126 and 60 exposed; one early and one late verdict are no-consensus). The agent sees the injected sentence *"Your assistant has identified the source cohort: TCGA [code] ([full name])"*, followed by notice that a sample-identifier mapping is available. The identity-verdict judge sees only the submitted subtype labels and mechanism hypothesis, together with the true and planted candidate identities. The support judge sees an extracted, agent-authored trace (`# WHY` comments, submitted observations, subtype labels, top genes and mechanism) plus a true-cohort reference card for fact-checking; it does not receive the system prompt, disclosure preamble or tool-result text containing the injected label. The premise-test judge sees the same extracted trace plus the supplied label and its position, and no reference card. Neither judge determines exposure; delivery is reconstructed from the episode log.

Among the 184 exposed episodes with an identity verdict, support is grounded for 104, unsupported for 45, anchored for 26, and has no panel consensus for 9. The preregistered binary contrast compares grounded with the combined non-grounded group per model (Table 4); the nine support-no-consensus episodes are excluded from that contrast, not silently assigned.

A.19 Artifacts detected by the analysis rules

Process metrics share a failure mode in which a non-event is rendered as a benign value. Each rule in E4 exists because one such non-event produced a clean headline in our own interim reports; three are recorded here, with three further checks on the instruments.

The harness is part of the measurement (a leaked identity channel). An early harness wrote each episode's working directory into a readable path containing the cohort name, and the agent reasoned from it in $16/126$ pilot episodes. The interim finding that derivation predicts robustness rested on this channel; with the path scrubbed ($0/168$ clean episodes) the contrast is a null ($p = 1.00$). Agents use whatever channel is left open.

Exposure is the denominator (an intervention that never fired). The planted label is delivered on the Nth recorded observation with no fallback. Lean episodes stop at a median of four observations, so in the late-reveal lean cell 56 of 64 episodes were never shown a label. Over the arm, the late reveal reads as 36.7% adoption against 76.2% early ($p = 1.8\text{e}{-10}$); among exposed episodes the two are 76.3% and 76.8% ($p = 1.0$). Non-exposure had been recorded as resistance. Conditioning on exposure repairs the denominator but selects longer runs, so timing and scaffold contrasts on this arm stay descriptive (Fig. 6).

Stratify before pooling (a hypothesis confirmed by composition). With four families in the ladder, the preregistered contrast between grounded and non-grounded identity claims reads 66.3% vs. 88.7% adoption on the exposed set ($p = 7\text{e}{-4}$), and the strategy contrast 50.0% vs. 89.9% ($p = 1\text{e}{-6}$). Both are Simpson's paradox (Simpson, 1951): Opus 5 is at once the most grounded model ($45/51$), the most explore-tagged and the only one that resists. Without Opus 5 the first is 96.6% vs. 92.3% ($p = 0.44$), and within every model both are null (§A.4). The preregistration fixed the reading in advance, as a null, and we follow it.

The identity channel that blinding cannot close. The sandbox removes sex and diagnosis, but a public cohort's sample count, death rate and age distribution remain recognisable. A tightened

probe, the correct cancer named within 120 characters of the exact sample size before the reveal, fires in $10/42$ of Opus 5's blinded episodes and 0–1 of 42 for each other model (§A.12). The model that resists the false label is also the one with the most channels to the true one.

Judging without the answer key. The support judge fact-checks against a card naming the true cohort. Rerun without it (§A.7), the support label is substantially stable (κ 0.65–0.67, public judges), the G0→G2 explore shift is larger ($+45 \rightarrow +64$ staged, $+36 \rightarrow +49$ lean), and grounding and every within-model C3 contrast are unchanged: the card was damping the shift, not producing it.

Instruments require known-value arms. The two provenance instruments disagree on $259/445$ jointly labelled episodes with one reverse disagreement; they are nested, not interchangeable (§A.9), and the G0 check separates them. The premise test's control is the same arm, where rejection is known to be wrong, and no model rejects there. The strategy tag is near ceiling on the mechanism (80–95% across prompt × arm cells): every mechanism is built on computed statistics and recall enters at the interpretation, which the tag does not read, so 109 of the 141 episodes that adopted a false label carry an explore-tagged mechanism for the wrong cancer. The tag is informative only on a commitment that can itself be recalled, hence the timestamped channel (Fig. 3b). Process labels are judge-dependent where outcome verdicts are far less so: the gate rate varies by 1.2 points across families while D2 explore counts are $401/102/222$, and the premise test's "tested" rate spans 41–64% across families for GPT-5.5 and Sonnet 5 while Opus 5's stays above 96% (§A.8, §A.10).

A.20 Applying the protocol to other tasks

Applying the protocol to a new task requires a commitment elicited from the agent, a harness that logs its own assertions, and a channel through which a premise can be planted; ground truth and a memorised prior were needed here to validate the instruments and are not needed to run them. Cohort identity on public genomic data has all of these, as do a dataset's provenance in a data-analysis benchmark, a compound's activity in a chemistry agent, and a theorem's status in a mathematics assistant. For anyone deploying such an agent, the present study suggests three practices.

1. Withhold the recallable commitment, and expect the mask to leak. Withholding was the one lever that moved the balance toward exploration, and only part of the way: most withheld identities were recovered from the vocabulary reveal, and one model recovered them from the shape of the dataset. Measure the channel rather than assuming it closed.
2. Treat every assertion the harness makes as an untrusted input. A plausible false identity through an authoritative channel was adopted in $126/133$ exposed episodes by three of four families, and agents annotate a validated partition with whatever the label evokes, so a rigorous write-up is where a wrong identity is most convincing. A filename, prior diagnosis or metadata field is the deployment analogue: verify it, and audit identity separately from the analysis.
3. Score the arbitration before trusting the agent, and do not infer it from the outcome score. Most models test a supplied label; what discriminates is whether the agent's evidence or the channel prevails when they disagree, and whether the same agent accepts true labels. A higher outcome score is evidence of neither.

Of the levers tested, the environment moved the balance, the prescribed procedure did not, and the model did. An explicit instruction to weigh computed evidence over a supplied label was not tested and is the first lever we would try.